Netze, Karten, Irrgärten:
Graphenbasierte explorative Ansätze zur Datenanalyse und Anwendungsentwicklung in den Geisteswissenschaften

Rasmus Krempel

Inauguraldissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät der Universität zu Köln
im Fach Historisch-Kulturwissenschaftliche Informationsverarbeitung
Vorgelegt von Rasmus Krempel
Datum der Defensio: 04.05.2016

1. Gutachter: Prof. Dr. Manfred Thaller
2. Gutachter: Prof. Dr. Reinhard Förtsch
2. Gutachter: Prof. Dr. Øyvind Eide

Dank allen, die mich im Prozess der Erstellung und des Abschlusses dieser Arbeit unterstützten.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Daten und Information werden weithin für die Triebfeder des Fortschritts gehalten. Dabei wird ihr Unterschied oft verkannt. Informationen sind dabei per Definition wichtig, da sie neu sind. Der Begriff lässt sich von der Redundanz abgrenzen, welches Bekanntes darstellt. Der Begriff der "Daten" hat dabei einen faden Beigeschmack. Daten sind keine Informationen, da sie nicht zwangsweise etwas Neues darstellen. Eine weitere Abgrenzung muss zum Rauschen gemacht werden, den Daten die keine Information enthalten. Bei Daten ist unbekannt, woran man ist. Daten werden oft in Verbindung mit Sperrigem und Unverständlichem gerückt. "Datenhalden", "Datenberge", "Datengebirge" sind nur einige der herablassenden Begriffe für unverstandene, potenziell wertlose Information oder redundante Daten. Bei Begriffen wie "Datenberg" handelt es sich um Daten, die vom Menschen nicht erfahrbar sind oder aufgrund der begrenzten Lebenszeit eines Menschen nicht vollkommen erfahren werden wollen. Gerade in Kombination mit dem Begriff "digital" wird der Eindruck eines wertlosen Haufens aus Nullen und Einsen verstärkt. Das ist jedoch falsch, da alle digital vorliegenden Daten, selbst wenn es sich um aufgezeichnetes zufälliges Rauschen handelt, einem menschengemachten Grundgedanken entspringen. Das Rauschen würde als Ausgabe eines Sensors genau der Ausgabelogik bzw. Ausgabeformats dieses Sensors entspringen.
Folgt man der Argumentation, dass wir in einer Aufmerksamkeitsökonomie[1] [230] leben, dann sind digitale Datenberge, die zu Fuß bestiegen werden sollen, die Metapher eines persönlichen Bankrotts. Im Hinterkopf werden die Stimmen der Digitalisierungskritiker laut, die kulturpessimistisch und medienkritisch eine "digitale Flut" voraussagen, in der wir zu ertrinken drohen. Dies bemerkt man heute nur noch, wenn die Stimmen dieser Kritiker nicht schon längst in der Informationsflut der digitalen Nachrichten untergegangen sind. Sie haben aber in dem Sinne ein richtiges Argument, dass Kulturtechniken und Umgangsformen mit diesen Datenbergen etabliert werden müssen.
Als Flut können Änderungen in den Formen der Medienverbreitung gesehen werden. Die

[1] Die Aufmerksamkeitsökonomie sagt vereinfacht aus, dass die Produzenten um die Aufmerksamkeit des Rezipienten werben müssen, um in der Masse der Werke wahrgenommen zu werden. Der Begriff ist stark an die Digitalisierung gebunden, da nun die Anzahl der verfügbaren Werke die mögliche Aufmerksamkeit stark überschreitet. Für den Konsumenten entsteht ein Auswahlproblem und für den Anbieter ein Problem überhaupt wahrgenommen zu werden.[230]

Medien ändern dabei nicht zwangsweise ihre Form. Video, Text, Bild und Audio werden zwar hypertextuell erweitert, doch im Grunde bleiben sie, was sie sind. Video hat sich auf digitale Onlineplattformen verlagert. Es kann "on demand" konsumiert werden, wo früher der Rundfunk Konsumzeiten vorgab. Gleiches gilt für Ton. Neben dem "Wann" hat sich auch das "Wo" verändert. Mobile Endgräte machen den Konsum insofern unabhängig vom Fernseher im Wohnzimmer. Bilder kommen aus dem persönlichen Fotoalbum ins Netz, wo sie potenziell jeder sehen könnte und wo man sie mit Freunden teilen kann. Text verlagert sich vom gedruckten Werk auf Webseiten und E-Books.
Der wichtigste Umbruch ist dabei der Zugang und das in zweierlei Hinsicht. Einerseits verschwimmt die Trennung von Rezipient und Produzent, was die potenziellen Datenproduzenten erhöht.
Andererseits ist der Zugang zu Informationen und Medien durch das Internet einfach. Die potenzielle Auswahl ist riesig, die Art der Auswahl wird dadurch zu einer Schlüsselkomponente bei der Bewältigung der digitalen Flut. Dabei sind die Daten die für das Auffinden verwendet werden nicht die Daten selber, sondern Daten über die vorhandenen Daten. Allgemein werden diese als Metadaten bezeichnet.
Der Mensch braucht Kulturtechniken, die das Mehr an Daten verwaltbar machen und in wirkliche Information verwandeln können, andere Techniken als ein ungefiltertes zusammenhangloses Abspielen des Inhalts.
Zur Datenschwemme aus dem Medienbereich kommen immer mehr Daten von Kameras, Mikrofonen, GPS Geräten und sonstigen Sensoren die im festen Takt Daten ausspucken. Bei Kameras gibt es Bemühungen zur automatischen Gesichtserkennung also dem Erkennen von Personen auf statischen und bewegten Bildern. Auch sollen Handlungen erkannt werden. Der Computer soll "intelligent" werden, sodass das Aufgezeichnete auch "verstanden" werden kann. Das ist notwendig, da der Mensch nicht mehr die Zeit hat, alle diese digitalen Eindrücke, die nach dem Vorbild seines sensorischen Apparats geschaffen wurden, zu verarbeiten.
Dabei ist der Mensch aufgrund seines Sinnsystems dazu in der Lage, große Datenmengen in Form von Videodaten zu verarbeiten. Wenn die reine Informationseinheit gesehen wird, dann kann der Mensch 100 Megabyte Video sehr viel schneller verarbeiten als die gleiche Datenmenge in Textform. Daher sind "Datenberge" nicht zwangsweise durch die reine Menge an belegtem Speicher zu definieren. Es geht dabei eher um den Zugang für den Menschen. Der Zugang zu Videodaten ist ohne Videosoftware nicht möglich. Videodaten sind in gewissermaßen ausschließlich maschinenlesbar. Die Maschine übersetzt sie in etwas, das der Mensch versteht, in Bild und Ton.
Prinzipiell kann der Mensch nur begrenzt Daten im Binärformat verstehen und selbst die Darstellung von Text benötigt Programme, doch ist die Abstraktion und die Datenmenge viel geringer. Des Weiteren werden Symbole verwendet wie Wörter, Buchstaben, Zahlen, welche durch die Eingabe klar reproduzierbar und auffindbar sind. Bei Gesichtern und Geräuschen ist die Klassifikation und Beschreibung nicht so einfach wie mit textbasierten Informationen.
Daraus entsteht das Problem, dass wieder Daten anfallen, die durchsucht, gesichtet und

überprüft werden müssen. Es handelt sich also in gewisser Weise um eine Übersetzung von Daten in ein Format, das mit bekannten Methoden durchsucht werden kann, Metadaten. Zum großflächig automatischen Erschließen neuer Daten kommen Digitalisierungsanstrengungen um Altdaten zu erschließen. So kann aus gescannten Dokumenten automatisch Text extrahiert werden. Digitalisierungsanstrengungen, bei denen ganze Bibliotheken gescannt wurden wie Google Books sind nur populäre und kontrovers diskutierte Beispiele der Digitalisierung im großen Stil. Nicht beachtet werden dabei die riesigen Archive aus Verwaltungsdaten, Akten und Katalogen. Im Bereich der Digital Humanities gab es im letzten Jahrzehnt eine starke Förderung von Digitalisierungsprojekten. Nach der Phase der Digitalisierung muss auch gefragt werden: Welcher Mehrwert kann aus diesen Daten gezogen werden und wie können diese sinnvoll verarbeitet werden? Das lineare Lesen all dieser Texte ist nicht möglich. Das Erfassen von Information auf anderem Wege wird wichtig.[62]

Auch wenn das Lesen auf einem E-Book-Reader möglich ist, handelt es sich dabei nicht um methodischen Fortschritt. Ein E-Book ist die digitale Imitation eines Buches. Es werden viel mehr Dinge viel einfacher verfügbar, die verfügbare Zeit diese zu verarbeiten ist dabei im besten Fall gleich geblieben.

Ein Fortschritt lässt sich bei maschinell verarbeitbarem Text dennoch finden. Digitale Texte haben den Vorteil, dass sie direkt auf Symbolebene verarbeitet werden können. Es müssen also keine Metadaten vorhanden sein, die den Inhalt abstrakt beschreiben.

Ein geistiger Einstiegspunkt in diese Ebene kommt aus der theologischen Forschung aus einer Zeit vor der Aufmerksamkeitsökonomie. Hier ist die Rede von Roberto Busa und dem Ursprungsmythos der Digital Humanities.[35] Bei Busa wurde der Computer zum Ordnen und dem lexikalischen Verarbeiten von Sprache eingesetzt. Aus seinen tausenden von Karteikarten wurde mit Hilfe von IBM der Index Thomaikus[47] geschaffen, ein Meilenstein der Informationstechnik. Auch wurde durch eine frühe Form von Hypertext ein Grundstein für das heutige Web gelegt.[241]

Anders als Bildscanns und Papierdokumente sind Textdokumente nach Buchstaben, Zahlen, Silben, Namen, Floskeln und Abschnitten etc. durchsuchbar. Das Auffinden von wichtigen Textstellen ist auf diese Weise einfacher geworden als beim Suchen mit Händen und Augen. Durch das Einbeziehen von Synonymen und Schlagworten wird das Auffinden erleichtert. Dabei beschränkt sich diese Lösung nicht auf eine Sammlung retrodigitalisierter Texte. Alle Daten, die auf Text beruhen, können so durchsucht werden. Von der Suchleiste im eigenen E-Mail Postfach bis zur Suchleiste eines großen Suchmaschinenanbieters, mit der das ganze Internet durchsucht werden kann, lässt sich dieses Werkzeug immer gleich bedienen. Eingabe ist eine Menge von Worten, die je nach Expertise des Nutzers logisch verknüpft, ausgeschlossen oder in Mehrwortgruppen zusammengefasst werden können. Suchleisten lassen sich ähnlich bedienen und liefern auch meistens das gleiche Ergebnis.

Die Ausgabe ist eigentlich immer eine nach “Relevanz” geordnete Liste von Dokumenten oder Positionen in Dokumenten. Dabei ist die Ordnung, die aus einem Gewichtungsprozess entsteht, nicht in jedem Fall nachvollziehbar. Bei großen Suchmaschinenbetreibern

liegt das daran, dass sie aus guten Gründen ihre Algorithmen nicht offen legen.[2] Auf der anderen Seite kann die Anpassung an den Nutzer besondere Suchergebnisse in der Relevanz aufsteigen lassen. Dieses Vorgehen wirkt dem Zweck einer Suche entgegen, potenziell Neues zu finden.
Auf der rein technischen Seite tragen eine Menge verschiedener Algorithmen zur Indizierung und der Vorverarbeitung bei.[3] Das macht, das Nachvollziehen des Zusammenhangs zwischen Eingabe und Ausgabe für den Laien schwierig.
Eine Volltextsuche ist dann vorteilhaft, wenn man weiß, was gesucht wird. Wenn jedoch größere Datenmengen erfasst werden sollen wie z.B. ein Katalog dann können Dinge gefunden werden, die vielleicht allein durch ihre Art uninteressant sind. Eine erweiterte Suchmaske könnte die einzelnen Sinneinheiten durchsuchbar machen. Die Strukturen und Muster, die in stark strukturierten Datensätzen vorhanden sind, bleiben dabei jedoch unbeachtet.
Ein Projekt, das einen quantitativen Zugang zu Millionen von gescannten Bücher ermöglichte, war das Google N-Gram Projekt. Es ging aus dem Google-Books-Projekt hervor. Bei Google N-Gram können die Bücher aus dem Google-Books-Projekt nach Phrasen einer bestimmten Länge (sogenannte N-Gramme) durchsucht werden. Diese Auftrittshäufigkeiten von Phrasen lassen sich nach dem Publikationsdatum des Buches und ihrer Auftrittshäufigkeit visualisieren.[110]
Haufen von Daten können natürlich mit der Volltextsuche nach Fragmenten durchsucht werden. Dabei stellt sich die Frage, ob es wirklich nur Haufen von Daten sind oder ob sich hier Strukturen finden lassen, die es ermöglichen neue und komplexere Fragen zu stellen. Es kann auch wichtig sein, ob die Strukturen von Daten aufzeigen, dass eine Frage mit einem Datenbestand nicht beantwortet werden kann. Es geht um Fragen auf einer höheren Ebene, die nicht allein mit dem Werkzeug der “Nadelstiche”, wie sie mit einfachen Suchen gemacht werden, beantwortet werden können!
Ein Mittel das im Onlinehandel, in Bibliotheken, bei Wikis und generell auf Datenbanken[118], eingesetzt werden kann, ist das Facetted Browsing[253]. Es erlaubt, in Kategoriesystemen dynamisch zu filtern. Die Auswahl von Kategorieausprägungen grenzt die Ergebnismenge immer weiter ein. Es wird also bis zu einem befriedigenden Ergebnis immer weiter gefiltert. Ein wichtiges Element, das den Verlauf dieses explorativen Vorgangs darstellt, ist die Breadcrump Navigation.
Die Breadcrump Navigation stellt den Filterprozess als Abfolge von Filtern dar. Sie werden in der Reihenfolge, in der der Nutzer Einschränkungen gemacht hat, abgebildet und ermöglichen so eine schnelle Orientierung des Nutzers.[150, 127]
Die anfangs beschriebenen Datenberge ziehen potenziell Datenberge nach sich. Daten über die Nutzung und Betrachtung der vorhandenen Datenberge. Die Kosten für Speicher nehmen ab und die Möglichkeiten Daten direkt aus digitalen Interaktionsprozessen wie dem Besuchen eines Online-Shops oder dem Benutzen eines Fahrkartenautomaten, zu gewin-

[2] Angesprochen wird dieses Problem in Abschnitt 2.2.3.

[3] Wichtige Vorverarbeitungen sind hier beispielsweise das tokenizing und stemming, aber auch das automatische einbeziehen von Synonymen oder auch ähnlich geschriebenen Worten.

nen, steigen.
Ein rein digitaler Vorgang ist das Analysieren von Nutzerverhalten beim Surfen im Internet. Das Zählen der Aufrufe von Websites ist ein schon lange verbreitetes Mittel Informationen über die Nutzung zu erhalten. Die Weiterentwicklung hierzu ist die Erfassung der Mausbewegung zu jedem Zeitpunkt. Die angeklickten Dinge werden mit dem Profil des Nutzers verbunden und gespeichert.
Ähnlich wird auch die Aufzeichnung der Augenbewegung[4] genutzt. Im Zuge der Aufzeichnung durch Kameras ist dies auch eine Möglichkeit, um potenzielle Konsumartikel des Begehrens einem Nutzer zu zuordnen. Verwertet werden diese Informationen in riesigen Zuordnungsnetzen von Personen zu Dingen. Diese Netze speisen Vorschlagssysteme. Vorschlagsysteme [5] zeigen dem Nutzer neue und vermeintlich begehrenswerte Dinge auf.[274, 276]

1.1 Ziel der Arbeit

Das Ziel dieser Arbeit bezieht sich auf die explorative Analyse von bestehenden, strukturierten Datenbeständen und auch in strukturierten, annotierten Texten. Dabei werden Fragen der Semantik und der Heterogenität von Netzen besprochen und welche Informationen daraus abgelesen werden können.
Die Stichworte, die das Vorgehen am besten beschreiben, sind Data-Mining, Knowledge discovery, (soziale) Netzwerkanalyse. Die Daten der Digital Humanities sind dabei stark an Text und Metadaten orientiert, von Haus aus wird weniger auf rein mathematische Ansätze zur Problemlösung gesetzt.
Zieldarstellung sind Netwerkdiagramme und Netzwerk basierte Zugänge zu Datensätzen. Diese sollen helfen, verständliche und problematische Muster in Daten zutage zu fördern.[93] Dabei kann man nicht mehr nur statische Darstellungen als Ziel einer Datenbetrachtung sehen. Interaktive Darstellungen und Zugänge zu Datensätzen sind heutzutage Standard. Dies ermöglicht es beispielsweise einen Rückschluss auf die Quelldaten zu bekommen wie beim Close Reading.[81] Im weiteren wird auch versucht die Vorteile einer automatischen Darstellung mit denen einer statischen Darstellung zu vereinen wie in Kapitel 5 gezeigt wird. Um in einer Welt mit zusammenwachsenden Datenbeständen interessante Phänomene extrahieren zu können, ist es im weiteren auch wichtig, Grenzen einer Betrachtung zu erkennen oder selbst zu setzen und zu interpretieren. Hier kommen gerade die Bedingungen und Fallstricke dieser Vernetzung zum Tragen. Denn um eine aussagekräftige Ansicht zu finden, muss ein Rahmen definiert werden. Auch lassen sich aus manchen extrahierten Netzen weitere Anwendungen bauen. Um Anwendungen auf der Grundlage von Daten zu bauen, muss erkannt werden, welches Potenzial in den Daten steckt.
Die Digital Humanities sind durch den englischen Begriff der Humanities geprägt, der sich nicht mit den Geisteswissenschaften deckt, sich aber doch von technischen und naturwis-

[4] Eyetracking

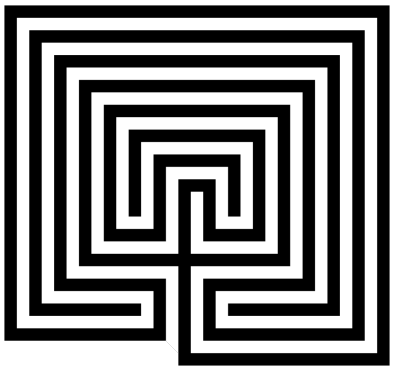
[5] Engl. recommender system.

senschaftlichen Fächern abgrenzt. Es dreht sich also um eine Betrachtungsweise aus einem Feld von Disziplinen, die nicht in erster Linie mathematisch fundiert arbeiten. Trotzdem muss in diesem speziellen Fach der Digital Humanities oder auch digitalen Geisteswissenschaften vorausgesetzt werden, dass eine Grundlage an informatischem Wissen vorhanden ist.
Es wird ein Verständnis für Informationstechnologie vorausgesetzt, welches vom Autor nicht zwangsweise mit dessen mathematischer Grundlage gleichgesetzt wird. Die Ausarbeitung dieser Arbeit versucht diesem Punkt Rechnung zu tragen, indem keine Flut an Formeln verwendet wird, sondern eine Vermittlung der Konzepte, deren Einschränkungen und Konsequenzen durch Visualisierungen, Beispiele und möglichst einfache Szenarien aufgezeigt wird. Die formale Ausdrucksweise, wie sie in der Informatik verwendet wird, ist hier zugunsten einer allgemein verständlichen Ausdrucksweise ausgelassen. Die mathematische Notation wird die meisten Geistes- und Kulturwissenschaftler fragend zurücklassen, obwohl die Konzepte hinter solchen Methoden nicht zwangsweise der Komplexität ihrer Notation entsprechen.
Auch wird der Modellierung von Ideen und deren Gestaltung ein höherer Wert beigemessen als einer zeitkritischen Verarbeitung. Auf der Ebene der Algorithmen werden daher eher intuitive Lösungen aufgezeigt, um einen Prozess zu illustrieren, in den einfach gestalterisch eingegriffen werden kann, so wie es auch das Ziel der meisten Programmiersprachen ist.
Durch das exponentielle Anwachsen des Speichers und der Rechenleistung[174] können viele Probleme auch ohne Optimierung gelöst werden. In der Praxis kann ein handelsüblicher Desktop-Computer auf einer mittelgroßen Datenmenge oft ausreichend schnell Ergebnisse liefern. Falls der Modellierende eine Lösung findet, die im weiteren optimierungswürdig ist, dann ist das Optimieren selbst die Aufgabe für einen Fachmann, das natürlich nur unter dem Gesichtspunkt, dass nicht schon eine Optimierung vorliegt, die nur erschlossen werden muss wie eine vorhandene Programmbibliothek.

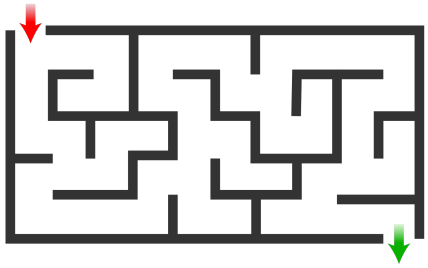
1.2 Netze, Karten, Irrgärten

Karten und Satellitenbilder, die einem den Einblick in komplexe Straßensysteme geben, bieten Orientierungshilfe im Alltag. Viele touristisch erschlossene Städte bieten Broschüren mit einer Karte darauf an oder einen teuren Souvenirshop mit Stadt- und Ansichtskarten. Ansonsten können auch Smartphones und Tablets mit Karten, on- oder offline, genutzt werden. Es ist normal, über eine Abbildung der Welt aus der Vogelperspektive zu verfügen. Wenn ein Solcher Einblick fehlt dann stellt das ein massives Problem für die Orientierung dar. Es gibt Beispiele wie Zugriffsbeschränkung durch das Verwehren dieses höheren Einblicks implementiert wurden.
Im Folgenden werden Computerspiele als Beispiele für eine Betrachtung des Zugangs zu Information besprochen. Dies geschieht aus folgenden Gründen:
Die Daten eines Computerspiels sind in sich geschlossen vollständig, es ist alles über sie bekannt und es lässt sich alles herausfinden.

Abbildung 1.1: *Vergleich von Labyrinthtypen*



(a) *Klassisches kretisches Labyrinth, ein verschlungener Weg ohne Abzweigungen und Sackgassen.[7]*



(b) *Ein Irrgarten mit Abzweigungen und Sackgasse, im Englischen Maze.[138]*

- Sie beinhalten oft komplexe Rahmenhandlungen, Interaktionen zwischen Charakteren und Ereignisse.
- Sie beinhalten oft geografische Komponenten.
- Sie lassen sich anders als die Realität komplett durchdringen und dokumentieren, Metainformationen können vollständig sein.

Netze gleichen häufig Irrgärten, wenn ihr Aufbau nicht immer hinreichend bekannt ist. Hier ist eine Abgrenzung vom klassischen Labyrinth[6] ohne Kreuzungen und tote Enden wichtig[7] siehe Abbildung 1.1. Der Irrgarten ist wie ein Spielplan, der Spielplan ist einem Spieler jedoch meist in Form einer Karte zugänglich. Bei Irrgärten ist die Aufdeckung Teil des Spiels. Denn wenn keine Karte vorhanden ist, kein Überblick existiert, sondern nur eine persönliche Perspektive, dann kann es schwierig werden, ein Ziel zu finden.
Der Hauptgrund ist: Der Weg zum Ziel ist nicht absehbar. Um den Überblick zu erlangen, kann man beim Erforschen eines Netzes alle Wege abgehen, irgendwann wird das Ende erreicht, solange man nicht im Kreis läuft und immer denselben toten Weg abgeht.
Abgesehen vom Spielfeld oder der Spielwelt, durch die sich der Avatar[8] des Spielers bewegt, gibt es auch andere Netze in Spielen, die nicht auf der Position von Spielfiguren oder Avataren beruhen. Computerspiele, die Geschichten erzählen, beinhalten auch Netze anderer Art. Sie können von der Handlungsstruktur Labyrinthe oder Irrgärten sein. Klassische Adventurespiele haben wie Bücher und Filme einen linearen Ablauf von möglichen Ereignissen. Es gibt interessante Ereignisse und Rätsel zu lösen, allerdings nur einen vorgegebenen, linearen Handlungsstrang. Diese Spiele sind wie Labyrinthe, durch ihre Windungen

[6]Im weiteren werden die Begriffe (klassisches) Labyrinth und Irrgarten verschieden unterschieden.

[7]Im Englischen ist die Unterscheidung einfacher mit dem "Labyrinth" als klassische Variante ohne alternative Pfade und dem "Maze", mit alternativen Pfaden.

[8]Als Avatar wird die digitale Repräsentation des Spielers in einer virtuellen Welt oder wie hier einem Computerspiel bezeichnet.

vermittelt sie den Eindruck als wäre mehr nötig, um es zu durchqueren als voranzuschreiten.
Auf der anderen Seite können Spiele komplexe offene Welten[9] sein, die zwar meistens Anfang und Ende kennen, dazwischen jedoch auf mehreren verschiedenen Wegen lösbar sind. Dies ist beispielsweise in neueren Rollen- und Adventurespielen zu finden. In diesen Spielen gibt es Handlungsalternativen, die den Ausgang eines Handlungsstrangs verändern. Dies ist bei Literatur und Filmen nur selten möglich. Hier gibt es Ausnahmen, ein breiter Erfolg hat sich jedoch nicht eingestellt. Wie in einem Irrgarten gehört es dazu, dass Handlungen und deren Konsequenzen nicht absehbar sind.

Der Weg durch den Irrgarten

Einfache Wege durch Irrgärten zu finden, ist eine Aufgabe, die von Computern gelöst werden kann. Auch Menschen lösen diese Art von Aufgaben in Spielen und Rätseln. Dabei ist die Lösungsstrategie für diese Rätsel interessant. Die "beste" Lösung ist nicht immer der kürzeste Weg, da es auch andere Anforderungen an einen Weg geben kann.
Jeden Weg abzugehe, ist eine sichere Möglichkeit, das Ende des Irrgartens zu erreichen. Das ist eine Sisyphusarbeit, eine einfache und zermürbend langwierige Aufgabe. Diese langwierigen zermürbenden Aufgaben lassen sich jedoch meist gut automatisiert lösen! Die einfache Aufgabe wird von einem Algorithmus beschrieben und die Ausführung wird dem Computer überlassen.
Eine Frage in diesem Zusammenhang ist: Wie "viel" Gedächtnis[10] wird gebraucht, um einen Pfad durch einen Irrgarten zu finden?
Sich alle beschrittenen Wege zu merken, nimmt mit der Menge der möglichen Wege zu. Dabei ist die Gedächtnisstategie wichtig, nicht jeder Weg hinter einer "toten" Abzweigung muss gespeichert werden nur der Abgang zu einem toten, abgeschlossenen Teil des Irrgartens.[11].
Besonders kritisch sind Kreise, sie machen es notwendig zu wissen, ob man an einer Stelle schon einmal war, ansonsten betritt man unbewusst einen Teil des Irrgartens, den man kennt, aber irrtümlich für einen ganz neuen, unbekannten Teil des Irrgartens hält. Algorithmisch ist dieser Teil des Problems als Speicherproblem zu beschreiben. Umso mehr man sich merken muss, umso mehr Platz braucht man im "Kurzzeitgedächtnis". Das gilt für den Menschen sowie den Computer.
Im Mythos um das Labyrinth des Minotaurus wird ein Faden verwendet, um beschrittene Wege zu markieren. Das Markieren der beschrittenen Wege im Netz ist auch eine gängige Praxis in Netzwerkalgorithmen. Der Vorteil eines solchen Modells ist die Externalisierung von Wissen in das schon existierende System. Die bessere Analogie ist in diesem Fall das

[9] In dieser Arbeit ist später von der Offene-Welt Hypothese zu lesen. Diese trifft gerade nicht auf Computerspiele zu, da Spiele abgeschlossene, voll erfassbare Systeme darstellen.

[10] Laut Miller kann das Kurzzeitgedächtnis eines Menschen etwa sieben Informationseinheiten (engl. chunks) gleichzeitig speichern.[171] Die genaue Zahl ist nicht ausschlaggebend, selbst kleine Netze bestehen aus mehr als sieben Kanten. Damit ist das Kurzzeitgedächtnis des Menschen bei einer formalen Bearbeitung mit hoher Wahrscheinlichkeit überfordert.

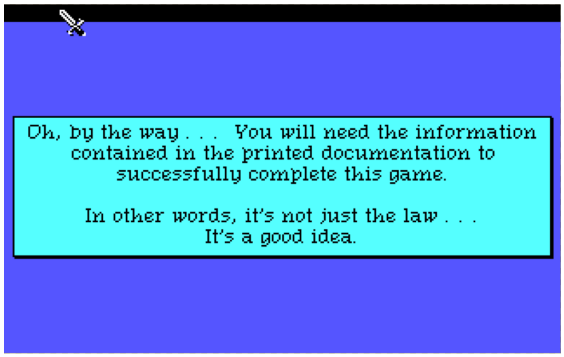
[11] In der Informatik "prunning" genannt. Äquivalent zum Beschneiden eines Baums oder einer Pflanze.

Markieren der Wände des Irrgartens mit Kreide, da Markierungen beliebig entfernt und hinzufügt werden könn. Ein “roter Faden” wird noch zusätzlich verwendet, um den schon beschrittenen Weg zu kennzeichnen. Dieses System ist vor allem wichtig, wenn die Umgebung an sich so gleichförmig ist, dass man Wege im Labyrinth nicht auseinanderhalten kann.
Der Mensch könnte eine Karte anlegen. Die Karte ordnet die geografische Information. Das reine Auslagern der Information würde in einer Liste der aneinandergrenzender Straßen ähneln. Zur Navigation müsste an jeder Kreuzung nachgeschaut werden, wo der nächste Schritt hingeht. Bei der Karte findet eine Ordnung der Informationen statt. Diese hilft, die nächsten Schritte zu planen, ohne immer die ganze Liste der Straßenverbindungen zu durchlaufen.
Für einen Computer entspricht dieses Vorgehen einer optimierten Datenstruktur im Speicher oder einem speziellen Index. Die Karte optimiert den Zugang für den Menschen auf visuelle Art, sodass sein Sinnessystem effizienter auf die Daten der Karte zugreifen kann.

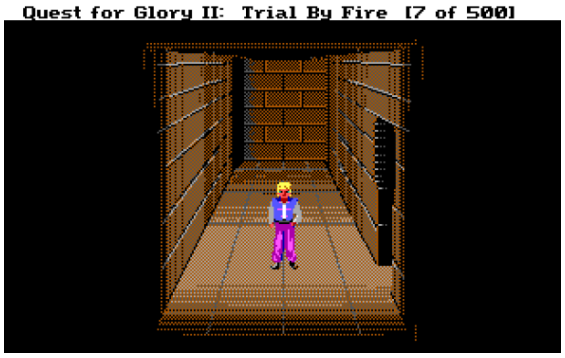
Der Kartenbesitzer im Irrgarten

Das Phänomen, dass sich Menschen ohne Karten schlecht orientieren können, wurde in der Computerspielbranche Ende der 80er, Anfang der 90er Jahre des letzten Jahrhunderts ausgenutzt. Hier dienten in einigen Spielen Karten, die bei einem Spiel mitgeliefert wurden, als Kopierschutzmaßnahmen. Mit den damals aktuellen Kopiergeräten waren die Karten nicht brauchbar zu reproduzieren, da sie farbig waren und die meisten Kopiergeräte nur Schwarz-Weiß-Kopien anfertigen konnten. Dieses Vorgehen wird im weiteren beispielhaft an den Spielen Starflight 2[29], Quest for Glory 2[87] sowie Grand theft Auto[70] angesprochen.
In den Starflight Spielen wurden Karten mitgeliefert, die ein Sternensystem darstellten, welches als Spielfläche dient. Im Spiel taucht bei den Reisen durchs Sternensystem zufällig die Sternenpolizei auf und erkundigt sich nach der Menge von Sternen einer bestimmten Farbe in einem Quadranten. Konnte der Spieler diese Frage nicht zufriedenstellend beantworten, wurde er aus dem Spiel geworfen. Anhand der Karte, die im Computer Grundlage des Spiels ist, können solche Informationen aus den Daten erstellt, aber auch vom Menschen anhand einer gedruckten Karte beantwortet werden.
Ein weiteres Beispiel für diesen Mechanismus ist das Action-Adventure-Rollenspiel Quest for Glory 2. Es spielt in einer fiktiven orientalischen Fantasy-Welt mit dem Thema von Tausend und einer Nacht. Ein Kopierschutz besteht daraus, im Straßensystem aus engen Gassen den Geldwechsler zu finden. Das ist eine Aufgabe direkt am Anfang des Spiels und notwendig, um weiter spielen zu können. Der Geldwechsler befindet sich dabei am Ende einer verzweigten und sich oft windenden Straße mitten im Spielfeld. Die Aufmachung der Straßen wie in Abbildung 1.2c gezeigt, ist generisch und lässt keine einfache Orientierung zu. Das machte es nahezu unmöglich den Avatar ohne Karte zum Ziel zu führen. Das ist keine technische Unzulänglichkeit, was sich daran zeigt, dass sich der Spieler mit dem gewechselten Geld im Spiel eine interaktive Karte kaufen kann. Diese Karte ersparte dem Spieler den ständigen Irrgang durch die Gassen. Des Weiteren weisen die Entwickler

Abbildung 1.2: *Beispiel des Kopierschutzes beim Quest for Glory 2*



(a) *Hinweis des Publishers zum Kopierschutz durch die Anleitung*



(b) *Innenansicht des Labyrinths*

(c) *Interaktive Karte im Spiel*

ausdrücklich auf einen Kopierschutz hin, wie in Abbildung 1.2a zu sehen ist. Im weiteren Sinne kann hier auch der erste Teil der Grand Theft Auto Reihe genannt werden, hier wurde auch eine gedruckte Karte zum Spiel mitgeliefert. Das Spiel dreht sich um das Stehlen von Autos und Verfolgungsjagden in Autos. Die Ansicht auf das Spielgeschehen ist eine dreidimensionale Ansicht, aus der Vogelperspektive über dem Avatar. Steigt der Spieler in ein Auto und fährt schnell, zoomt die Ansicht herraus. In Missionen wird man zwar per Pfeil in die Richtung der Missionsziele geführt, längere Distanzen oder Hindernisse wie Flüsse können dabei durch den Sichtradius des Spielers nicht erfasst werden. So wird es schwierig, sich bei längeren Strecken zu orientieren und nicht an Flüssen oder Sackgassen zu scheitern.[12]

Diese Sicherheitsmaßnahmen kann man die Klasse der Security through Obscurity[13] Methoden sehen. Für die Entwickler ist die Verwendung und Erstellung von Karten von Spielwelten einfach, da die Spielwelt einfach nur kleiner abgebildet werden muss. Die Daten sind verfügbar, richtig und vollständig. Dabei liegt die Sicherheit darin, dass der unberechtigte Nutzer aufgrund von fehlendem Hintergrundwissen und einer fehlenden intuitiven Handlung keinen Missbrauch durchführen kann.

Dabei treten in allen drei Fällen verschiedene Eigenschaften von Karten in den Vorder-

[12] Die Karten des Spiels sind auf der Website des Publishers abrufbar: `http://www.rockstargames.com/gta/cities/libertycity.html`

[13] Sicherheit durch ein unverständliches und undokumentiertes System.

grund:
Im ersten Beispiel zeigt sich, dass eine visuelle Erschließung von Informationen für den Menschen anhand von Karten einfach umzusetzen ist, die auf reiner Textbasis aber nicht so einfach möglich ist.
Das zweite Beispiel zeigt, dass Zusammenhänge im Nahen nicht gut erfassbar sind. Ihre große und sinnhafte Struktur kann auf diese Weise nur schwer erfasst werden.
Das dritte Beispiel demonstriert, wie eine umfassende Karte auch bei einer schon bestehenden Aufsicht ein nützliches Mittel zur Orientierung sein kann.
Später wurden diese gedruckten Karten immer mehr zu sogenannten "Goodies", welche den Spielen als Kaufanreiz beigelegt wurden. Dass Karten nicht mehr als "Kopierschutz" verwendet werden, kann auf den Fortschritt in der analogen Druck- und Kopiertechnik zurückgeführt werden. Auch stiegen mit der Entwicklung von interaktiven Online-Karten die Ansprüche an Computerspiele stetig. Schon seit Jahren gehören auf Knopfdruck verfügbare, interaktive Karten in Computerspielen zur Grundausstattung. Diese beinhalten viele Eigenschaften, die auch für reale Kartenanwendungen Standard sind wie Positionsanzeige für den Avatar, der Markierungen von besonderen Orten und Aufgaben. Dabei kommen gerade 3D Spiele mit großen interaktiven Spielwelten nicht ohne solche Karten aus.

Ohne Start und Ziel

Die Beschreibung des Netzes als Irrgarten erweckt den Eindruck, dass es sich um eine Struktur mit festem Anfangs- und Endpunkt handelt. In einer umfassenden Beobachtung ist jeder Punkt ein potenzieller Anfangs- und Endpunkt. Die persönliche Reise durch den Irrgarten tritt damit in den Hintergrund. Es wird kein Weg durch den Irrgarten gesucht, es wird versucht zu beschreiben, welche Eigenschaften der Irrgarten hat und wie er "aussieht". Wenn die Analogie des Irrgartens als persönliches Irrlaufen verlassen wird, dann verändern sich die Fragen, die an Netze gestellt werden.
Für die Bewertung von großen wie kleinen Netzen gibt es algorithmische Bewertungsverfahren, welche die Wichtigkeit einzelner Knoten und Kanten errechnen können. Hier wird die Fragestellung wichtig, die an ein Netz gestellt wird. Für die Frage nach der Wichtigkeit in einem Netz anhand seiner Position anderen gegenüber gibt es aus der Social Network-Analysis Zentralitätsmaße. [14] Mit den Zentralitätsmaßen kann beispielsweise die Wichtigkeit von Akteuren anhand ihrer potenziellen Informationskontrolle in einem Kommunikationsnetz aus Akteuren beschrieben werden.
In einem Rollenspiel, in dem der Avatar immer kleine Aufgaben lösen muss, wäre dies ein Ort, der in jeder Spielart immer wieder besucht oder durchquert werden muss, um ein Rätsel zu lösen oder eine Aufgabe zu erfüllen. Als Beispiel würde hier ein Gasthaus stehen, in dem der Avatar andere Nicht-Spieler-Charaktere trifft oder während seiner Reise ausruhen muss.
Ein anderes Beispiel ist die strategische Position von Brücken. Auf der einen Seite ist eine

[14] Siehe Abschnitt 2.5.4.

Brücke nicht viel mehr als ein meist schmaler Weg zwischen zwei Punkten. Doch ist es mehr als offensichtlich, dass Brücken eine einzigartige, zentrale Position haben. Sie verbinden zwei Punkte, die nur mühsam über andere Routen erreicht werden können.
Die Betrachtung eines Netzes kann aus einer Menge einfacher Fakten bestehen. Lokale Handlungen Einzelner und ihrer Beziehungen untereinander erzeugen Dynamiken, die im Kleinen für sich betrachtet nicht erkenn- oder erklärbar sind. Eine globale Sicht auf diese Einzelfakten wie Handlungen und Beziehungen kann einen Überblick über ein System verschaffen.
Für einen Spieler ist das Spielfeld die Menge der Möglichkeiten, seinen Weg durch ein Spiel von Start- zum Endpunkt zu finden. Hier zeigt sich auch der Unterschied zwischen dem einzelnen Spiel und dem Spielfeld. Das einzelne Spiel ist ein linearer Ablauf von Ereignissen, während das Spielfeld eine Auflistung aller möglichen Spielzüge ist. Ein unbekanntes Spielfeld oder ein Irrgartenspielfeld kann durch die Zusammenführung aus den Informationen aller Spiele und deren Spielzüge sichtbar gemacht werden. Dadurch kann der Spieler am Ende besser spielen, da er die Umgebung kennt.
Der Computer ermöglicht dabei eine effiziente Bearbeitung und Bewertung von Netzen, die methodisch vergleichbare Ergebnisse liefern können, wenn sie denn richtig eingesetzt werden. Per Hand die meist riesigen Datenmengen zu verarbeiten und den Überblick nicht zu verlieren, ist ein hoffnungslos anmutendes Unterfangen, so wie den Weg durch einen Irrgarten ohne Himmelsrichtungen zu finden, der nicht den Regeln des euklidischen Raums folgt.

Das Ende der Analogie

Die Irrgartenanalogie ignoriert, dass ein Netz nicht zwangsweise geopartialen Regeln folgt bzw. planar sein muss. Planarität lässt sich kurz als die Abbildbarkeit eines Graphen in zwei Dimensionen ohne Kantenüberschneidungen beschreiben.
In einem System, das per Definition planaren Regeln folgt könnten sich Kanten nicht kreuzen. Irrgärten sind meistens planar, sie lassen sich auf den klassischen zwei Dimensionen einer Karte abbilden. In einem realen Irrgarten kann noch die dritte Dimension hinzukommen. Durch Unter- und Überführungen können sich Wege überschneiden ohne das eine Kreuzung entsteht.
Eine Kante in einem theoretisch beschriebenen Graphen muss nicht den Regeln eines Weges in einem dreidimensionalen Raum folgen. Irrgärten befinden sich also in einem Raum und unterliegen den Regeln dieses Raums. In einem abstrakten Netz gibt es diese Beschränkung nicht.[261, Kapitel 2] Diese Probleme treten auf, wenn keine Ortsysteme, sondern Systeme von Beziehungen und anderen Verbindungen betrachtet werden. [15]

[15] eine Genauere Diskussion findet sich in Abschnitt 2.4.2.

1.3 Aufbau

Kapitel 2 zeigt im ersten Teil die gedanklichen Wurzeln netzwerktheoretischer Überlegungen aus verschiedenen Gebieten wie sie für eine im weitesten Sinne geisteswissenschaftliche Fachrichtung interessant sind. Anfangs werden in 2.2.1 Beispiele von Netzen betrachtet, wie sie in der sozialwissenschaftlichen Tradition verwendet werden. Weitergehend werden nicht nur die Möglichkeiten der Beziehung zwischen Menschen und Gruppen untersucht, sondern auch Beziehungen über und zwischen Werken und Dingen aufgezeigt.
Im Abschnitt 2.2.3 werden Informationsnetze besprochen und u.a. der Page Rank Algorithmus vorgestellt. Am Ende dieses Abschnitts stehen soziale Netzwerke. In Abschnitt 2.2.4 geht es um die Netze, die sich durch Wege und Straßen bilden. Auch werden Spielfelder als Handlungsspielraum betrachtet.
In Abschnitt 2.2.5 werden Anwendung auf Netzstrukturen besprochen.
Im zweiten Teil von Kapitel 2 werden grundlegende Modell- und Abbildungseigenschaften sowie Bewertungs- und Klassifikationsverfahren vorgestellt. Auch werden Aspekte der Netzwerkvisualisierung und der Wiederverwendung von Netzwerken besprochen.
Im Abschnitt 2.3 werden die Modellierungsmöglichkeiten von Netzen beschrieben und wie sich diese auf die Verarbeitungsmöglichkeiten auswirken. Des Weiteren werden Möglichkeiten der Visualisierung von Modelleigenschaften von Netzen angerissen.
In Abschnitt 2.4 werden einige Darstellungsweisen für Netze besprochen.
Im Abschnitt 2.5 Bewertungs- und Merkmalsbeschreibungen wie Grad, Dichte und Komponenten und die Zentralistätsmaße beschrieben, aber auch Clusteringverfahren und Verteilungs- und Konstruktionsphenomene wie die Long Tail Verteilung.
Abschnitt 2.6 zeigt verschiedene Repräsentationen von Graphen auf dem Computer auf, vor allem Austauschformate mit Fokus auf Ausdruckstärke und technologischer Grundlage.
Im Kapitel 3 Daten und Netze werden anfangs Datenstrukturen und deren Potenzial für die Netzwerkanalyse beschrieben. Dabei liegt der Fokus auf einzelnen Datenquellen, die für sich stehen. Dies sind Texte, Listen, XML und Geodaten.
Konkret demonstriert wird dies mit der digital ausgezeichneten Versionen von Macbeth und Gesetzbüchern, im speziellen des BGB. Im Abschnitt 3.3 wird ein Netz aus kontinuierlichen Wahldaten extrahiert. Dafür wird die Umwandlung von zahlenbasierten Wahldaten in relationale Daten demonstriert und Geodaten werden mit in die Darstellung eingebunden.
Der Abschnitt 3.4.1 Datenbanken gibt einen kurzen Einblick in Datenbanken, ihr Verhältnis zu Hypertext und dynamischen Inhalten im Internet und deren Verarbeitungspotential.
Im Abschnitt 3.5 wird ein parallelisier- und modifizierbarer Algorithmus skizziert, der Two-Mode Netze in One-Mode Netze projiziert. Im Anschluss wird dieses Vorgehen stukturiert auf die archäologische Bild- und Objektdatenbank Arachne angewendet und einige Ergebnisse diskutiert.
Das Kapitel 4 beschäftigt sich mit den Konzepten von Linked Data und dem Semantic Web und der damit einhergehen Entgrenzung von Daten. In Abschnitt 4.2 wird gezeigt wie semantisch verknüpfte Daten verarbeitet werden. Dafür wird, in Abschnitt 4.2.2, DBPedia

verwendet. Abschnitt 4.2.4 diskutiert die Bereitstellung von Daten und die Auswirkungen auf die Verwendbarkeit.
Abschnitt 4.3 zeigt praktische Beispiele für das explorative Vordringen in DBpedia auf Grundlage von Netzwerkvisualisierung. Dabei wird versucht auch problematische Stellen in den Daten aufzuzeigen.
In Kapitel 5 wird eine Betrachtung der in den vorherigen Kapiteln gesammelten Erfahrungen gerade im Hinblick auf dynamische Visualisierungen und exploratives Entdecken von Daten gelegt.
In Abschnitt 5.1 werden die Eigenschaften einer dynamischen Darstellung und die Möglichkeiten von Eigenschaften einer nachhaltigen und nachvollziehbar rekonstruierbaren Visualisierungsinfrastruktur dargestellt. Abschnitt 5.2.2 zeigt einige existierende interaktive Visualisierungen für Wikipedia und DBpedia. Abschnitt 5.3 beschreibt den Local Wikipedia Map Service der aus den Erfahrungen der Abschnitte 5.1 und 5.2.2 entwickelt wurde. Abschnitt 5.4 zeigt die Auswertung der Nutzung des LWMap Service. Am Ende in Abschnitt 5.5, werden die Erfahrungen aus dieser Arbeit zusammengefasst und ein Ausblick auf weitere Auswertungs- und Nutzungsszenarien gegeben.

Kapitel 2

Netzbetrachtungen

In diesem Kapitel wird im ersten Teil ein Überblick über die Anwendung von Netzen in den Geisteswissenschaften und ihnen nahestehenden Disziplinen gegeben. Im zweiten Teil werden die Modellierungsmöglichkeiten von Netzen beschrieben und welche Arten von Auswertungen sich daran vornehmen lassen. Des Weiteren werden Datenformate zur Abbildung und Programme zur Verarbeitung und Auswertung vorgestellt.

2.1 Übersicht

Die Welt spricht von global vernetztem Handel, von vernetzter Gesellschaft, dem Internet als weltweitem Netz. Soziale Netzwerke geben eine Gewissheit darüber, dass alles mit allem irgendwie vernetzt und vor allem erreichbar ist.
Die Betrachtung von Wissen als Netz zeigt ein abstraktes "Ganzes". Ein Netz, dessen Ausmaße nicht efassbar sind. Jeder Einzelne nimmt dabei eine oder mehrere Positionen in diesem Netz ein.
Der Einzelne sitzt an einer Stelle in globalen Wirtschaftsnetzen, Freundschaftsnetzen, dem Internet, etc. Dabei ist höchstens einen Teil des Netzes, im direkten Umfeld, sichtbar . Der Rest ist ein System, das sich der direkten Erfahrung entzieht.
Wenn die global vernetzte Ökonomie leidet, dann ist das eine abstrakte Bedrohung, die sich nur bedingt im eigenen Umfeld fühlen lässt. Der Einzelne ist nicht direkt betroffen. Er kann jedoch unverschuldet von etwas Abstraktem wie den falschen Risikoabschätzungen der Banken in fernen Ländern in seiner Existenz bedroht werden. Das geschieht, ohne direkt sichtbar zu sein. Der Pfad oder die Pfade der Konsequenzen werden erst dann sichtbar, wenn das Problem auftritt und ausgiebiger Prüfung unterzogen wurde.
Die Kosten des Internets erfährt der Einzelne nur durch die Kosten seines eigenen Internetanschlusses, für den ein Entgelt entrichtet wird. Es führt ein Kabel in die Wand, doch es steht ein riesiges globales Netz dahinter.
Hier muss nicht zwangsweise das Internet als Beispiel dienen. In gleicher Weise funktioniert ein interkontinentales Telefongespräch oder ein Brief der an das andere Ende der Welt gesendet wird. Das alles beruht auf solchen Systemen aus unsichtbaren Kommunikations- und Weiterleitungsknotenpunkten.
Netze bilden sich jedoch nicht nur im konkret Greifbaren, sondern auch in tradierten Ge-

danken als Produkt eines Netzes von Wissen, das zwischen Akteuren ausgetauscht und erweitert wird. Selbst ein Text wie dieser folgt lange tradierten Zeichen und Zahlensymbolen. Sie haben verschiedene Herkunft, die lateinischen Buchstaben und die arabischen Zahlen. Diese Gedanken liefen als immaterielle Konzepte durch die Köpfe verschiedenster Menschen und manifestierten sich in ihren Werken, bis sie zu einer Selbstverständlichkeit wurden. Diese Selbstverständlichkeit verschleiert die Komplexität ihrer Herkunft.
Dabei lassen sich Gedanken nicht immer ordnen und in eine klare Abfolge bringen. Definitionen lassen sich nicht aufstellen, ohne dass andere Begriffe benutzt werden. Diese müssen aber selbst definiert werden und das am besten, ohne den Begriff zu verwenden, den sie definieren sollen.
Netze, in der mathematischen Tradition "Graphen" genannt, lassen sich durch Knoten und Kanten beschreiben. Ein Netz in der realen Welt kann, je nach Betrachtung, sehr viele Knoten und Kanten enthalten. Die modellhaften Annahmen über das Verhalten und die Bedeutung von Knoten und Kanten sind von Fall zu Fall unterschiedlich. Es braucht eine Theorie, um zu erklären, was zwischen den Knoten über die Kanten passiert und warum nicht nur die direkt verbundenen Knoten wichtig sind, sondern ebenfalls die mittelbar verbundenen Knoten sowie die Struktur im Ganzen.

2.2 Netzwerk von ...

Netze gibt es viele verschiedene:
Handelsnetze, Karrierenetzwerke, Stromnetze, soziale Netze, Computernetze, Telefonnetze, Handynetze, Terrornetze, etc. Diese Liste ließe sich beliebig lange fortführen. Warum sollte eine bestimmte Theorie auf alle diese Sachverhalte passen?
Die Antwort ist ein gemeinsames Modell, mit dem alles abgebildet werden kann. Mathematisch kann alles als Netz dargestellt werden. Eine Menge von **Knoten** ist durch eine Menge von **Kanten** verbunden. Diese Verbindungen formen komplexe Strukturen. Das Bestehen oder Fehlen von Verbindungen zwischen zwei Knoten hat nicht nur Auswirkungen auf die betroffenen Knoten, sondern auch auf das Gesamtsystem. Indirekte Verbindungen über Dritte sind ebenfalls wichtig.
Hier gibt es Start-, End- und Übergangsknoten. Dadurch können Transaktionen oder Nachrichten auf verschiedenen Wegen durch das Netzwerk laufen.
Eine Verbindung muss dafür **transitiv** sein, das heißt, der Einfluss einer Kante geht über die an sie angeschlossenen Knoten hinaus. Der Anschluss der Nachbarschaft hat also zusätzlichen Wert.
Wenn ein Akteur als Knoten im Netz nur zwei Verbindungen hat, sagt das daher nichts über seine Fähigkeit aus, andere mittelbar zu beeinflussen. Er kann trotzdem potenziell Nachrichten, Handelswaren, Stromkapazitäten o.ä. kontrollieren beispielsweise, wenn er an einer exklusiven Position sitzt, die nicht umgangen werden kann. Diese Brückenpunkte oder auch Cutting-Points, können wichtige Rollen spielen. Ihre Position im Netz macht sie unersetzbar.
Manchmal ist auch nicht genau bekannt, warum Dinge verbunden sind. Websites können durch einen Link verbunden sein, von der einen in wie die andere Richtung. Doch **warum**

das eine Dokument auf das andere zeigt, muss im Zweifelsfall erlesen werden und die Erfassung hat im nackten mathematischen Modell des Netzwerks keinen Platz.
Ein klarer Fall sind wissenschaftliche Artikel und ihre Referenzen. Hier bedeutet eine Kante, dass ich auf einen Vorgänger bezogen wird. So müssten eigentlich Bäume[1] entstehen, da nur zitiert werden kann, was es schon gibt. Doch handelt es sich bei einem Baum auch um ein simples Netz mit strukturellen Einschränkungen. Aber das Modell der Knoten (Texte) und Kanten (Zitate) kann auch hier verwendet werden.
Die Anwendungen des theoretischen Konstrukts "Netz" sind also zahlreich. Es gibt sie fast überall. Gemein ist den meisten, dass die Systeme aus der Nah-Perspektive nicht sichtbar sind oder erst bei größerem Abstand ihre Gesamtdynamik zeigen.

2.2.1 Der Mensch verstrickt im Netz der Beziehungen

Der Mensch befindet sich am Anfang der Netzwerk-Analyse, wie sie als Grundlage für diese Arbeit genommen wird. So ist sie die Grundlage der sozialen Netzwerk Analyse, kurz SNA. SNA untersucht Beziehungen und Interaktionen zwischen Menschen, Gruppen und Organisationen.
Der Mensch tauscht Nachrichten aus, trifft sich, interagiert, steht in mehr oder weniger komplexen Rollen und Funktionen zueinander, übt Macht oder Befehlsgewalt aus etc.
Bei Gruppen und Organisationen sieht es ähnlich aus, nur dass es hier Abkommen, Verträge, Kooperation, Personen etc. betrachtet werden. Diese Arten von Beziehungen sind potenziell expliziter erfasst als rein menschliche Interaktion. In manchen Fällen sind diese Interaktionen mit Verträgen, Pressemeldungen etc. öffentlich zugänglich dokumentiert.
Diese Beziehungen und Interaktionen zwischen Menschen, Gruppen oder Organisationen werden durch Kanten abgebildet. Dabei können gerade die menschlichen Beziehungen sehr komplex sein. Genauso kann die Art von Kanten und deren Bedeutung variieren.

Der Einzelne und das ganze Netzwerk

Auf der Suche nach Netzen sind Ansatzpunkte nötig. Die Auswahl des Kriteriums, wer in einem Netzwerk ist und wer nicht in das Modell passt, muss definiert werden. In der Soziologie gibt es hierfür eine Trennung in der Methodik. Entweder es wird ein **soziozentrisches** oder **egozentrisches** Netzwerk betrachtet.
In **soziozentrischen Netzwerken** werden Personen als Knoten nach ihren sozialen Rollen oder über die Zugehörigkeit zu einer Gruppe ausgewählt. Diese Personen-Knoten werden über Kanten, die Beziehungen oder Interaktionen untereinander darstellen, verbunden. Hier werden über den Betrachtungshorizont hinausgehende Beziehungen nicht berücksichtigt. Die Kanten bilden die um eine Theorie aufgebauten Beziehungen zwischen den Personen ab. Die Schüler einer Schulklasse[141, 124] können als soziozentrisches Netz untersucht werden.
Ein Kritikpunkt an einem solchen Vorgehen besteht darin, dass bestimmte Gruppen in

[1] Baum im Sinne der Informatik.

soziozentrischen Netzwerken auch aus praktischen Gründen und weniger von einer theoriegetriebenen Fragestellung ausgehend gewählt werden. Eine Schulklasse stellt eine überschaubare und erfassbare soziale Einheit dar. Die Kinder in einer untersuchten Schulklasse haben aber je nach Lebensabschnitt auch andere wichtige Beziehungen außerhalb ihrer Schulklasse. Friemel und Knecht schlagen hier eine Differenzierung in ein **Primär-** und ein **Sekundärnetzwerk** vor, wobei das Primärnetzwerk den Untersuchungsraum die Klasse und das Sekundärnetzwerk alle anderen signifikanten Bezugspersonen der Schüler behandelt. [98]

Egozentrische Netzwerke[2] behandeln das Umfeld einzelner Personen und Organisationen. Als Ausgangspunkt wird ein Akteur genommen, der als "**Ego**" bezeichnet wird. Von diesem Ego ausgehend wird versucht die Beziehung von Ego zu anderen zu erfragen. Diese anderen werden "**Alteri**"[3] genannt. Bei diesem Verfahren gibt es Einschränkungen. Wer dazugehört das wird meist durch die aus der Fragestellung hervorgehenden Beziehungstypen bestimmt. Im soziologischen Instrumentarium ist auch ein Fragenkatalog zu Beziehungstypen und den Eigenschaften der "Alteri" vorgesehen. [267, 120]

In der soziologischen Methodik spielt die Beobachtung eine wichtige Rolle. Menschen verhalten sich anders, wenn sie wissen, dass sie beobachtet werden. Die soziologische Umfrageforschung und auch die Psychologie versuchen nicht direkt zu fragen, um ehrlichere Antworten der Probanden zu erhalten. Es wird versucht, indirekt zu messen. Es gibt also ein besonderes Instrumentarium um diese Daten zu erheben.

Bei der Erhebung von egozentrierten Netzen wird beispielsweise nach Personen gefragt, mit denen Ego bestimmte Aktivitäten unternimmt. Auf diese Weise soll indirekt gemessen werden, ob Personen befreundet sind. Diese Messung hat dabei eine andere Qualität als die direkte Frage nach Freunden. Wenn direkt nach Freunden gefragt wird, dann verzerrt die Sicht des Individuums auf sich selbst und sein Begriff von Freundschaft die Messung. Es werden daher andere Freunde genannt als die Personen, mit denen Ego "wirklich" nach der externen Definition, befreundet ist.

Anzumerken wäre, dass die meisten Wissenschaften auf diese indirekten Beobachtungen angewiesen sind, weil sie niemanden befragen können. Die Sozialforschung tut das jedoch ohne den Zwang, nicht direkt Fragen zu können.

Freundschaft

Ein Beziehungstyp, der wie erwähnt in der SNA untersucht wird, sind Freundschaften. Da Freundschaft schwer zu definieren ist, werden die meisten Sozialwissenschaftler mit einer praktischen Definition von Freundschaft beginnen. Freundschaft bringt eine Menge von Interaktionen zwischen Personen mit sich. Diese lassen sich messen. Freundschaften sind per se interessant. Typische Fragen sind: "Warum" sind Menschen befreundet? Wie entwickeln sich Freundschaften? Welchen Einfluss hat ein Freund oder der Freund eines Freundes auf eine Person? Befreunden sich Menschen eher mit Menschen, die ihnen ähnlich sind? Wie

[2] auch engl. Egonetworks oder personal networks.

[3] Engl.: alters.

entwickeln sich Netzwerke von Freunden? Gibt es Normen, die nur in bestimmten Freundschaftsnetzen oder Cliquen gelten? Dies sind einige Fragen, wie sie in der Soziologie gestellt werden.[140]
Eine Fragestellung, die dabei untersucht wird, ist der Einfluss von Freundschaftsnetzwerken auf das Verhalten von Einzelnen. Freundschaften scheinen zwischen Menschen, die sich ähnlich sind, also ähnliche Einstellungen haben und Werte und Normen teilen, persistenter zu sein.[140] Korrelation von Einstellungen geben dabei einen Ansatz zum Aufbau eines Graphenmodells.

Small-World

Eine Betrachtungsweise von einem "höheren" Punkt auf das menschliche Beziehungsnetzwerk ist das Small-World Experiment. Die Grundannahme ist: Über 6 Andere kann eine Person alle Menschen auf der Welt erreichen[4]. Diese 6 anderen werden dabei auch als maximale Distanz beschrieben die zwischen zwei Menschen existiert.[170]
Experimentell wurde dies von Travers und Milgram in den USA untersucht.
In ihren Experimenten wurde untersucht, über wie viele Stellen ein Paket weitergegeben wurde, bevor es den Empfänger erreicht. Das Paket durfte nur über Bekannte weitergegeben werden. Bekannte wurden nur als solche eingestuft, wenn die Personen sich gegenseitig kannten und beim Vornamen nanten. Diese Definition schließt berühmte Persönlichkeiten, die den Probanden nicht persönlich kennen, aus. Die Instrumentalisierung der Netzwerktheorie besteht hier darin, dass der Proband jemanden auswählt, den er als Verbindung zu jemand anderem in der Nähe des Ziels ansieht. Das Netzwerk der Bekanntschaften diente auf diese Weise als Transportsystem für das Paket.[252]
Diese Small-World Betrachtung ist dabei immer wieder auf andere Daten und Problemstellungen angewandt worden. Diese Entfernung hilft dabei, die Dichte eines Netzes zu beschreiben. Des Weiteren zeigt es, dass sehr kurze Pfade zwischen diesen existieren. Angewandt werden solche Informationen zum Erkennen von Dynamiken bei der Ausbreitung von Seuchen.[261, Abschnitt 2.1.1]

Der Feind meines Feindes

Auch in Geschichte und Literatur gibt es die Fragestellung nach übergreifenden Beziehungen: Kann der Feind meines Feindes mein Freund sein? Anders wäre der Freund meines Freundes auch mein Freund?
Diese Annahmen basieren darauf, dass es Widersprüche in Freundeskreisen oder Bündnissen gibt. Wie kann man mit Personen befreundet sein, die untereinander verfeindet sind? Diese Freund- und Feindschaftskombinationen lassen die Akteure in einem dissonanten Zustand zurück. Ihr Handeln Anderen gegenüber wird schwieriger. Allgemein formuliert wurde dieses Problem in der **Balance Theorie**.[129] Sie besagt, dass die nicht stabilen Zustände wie die der Freundschaft zum Freund eines Feindes oder der Feindschaft dreier Parteien untereinander zu einem ausbalancierten Zustand streben. Aus den unbalancierten Zuständen

[4] Engl.: six degrees of separation

werden sich balancierte Zustände entwickeln, in denen alle Dreiecksbeziehungen ausbalanciert sind. Dieser ausbalancierte Zustand kann durch das Wegfallen von Beziehungen erreicht werden oder durch das Ändern von Beziehungen von Freundschaften zu Feindschaften und umgekehrt.
Für das beschriebene Problem kann ein Netzwerk aus Freundschaften bzw. Bündnissen und Feindschaften als Graph mit positiven und negativen Beziehungen stehen. Zum Netzproblem wird die Balance-Theorie, da sich problematische Beziehungen nicht nur in einzelnen Dreiecken wahrnehmen lassen sondern auch in komplexeren Netzen aus mehr als drei Knoten. Hier kann es zum Problem werden, dass eine Beziehung in einem Dreieck ausgeglichen ist und im anderen Dreieck eine Dissonanz hervorruft.
Rechnerisch gibt es für ein System, das den Regeln der Balance-Theorie folgt, nur zwei stabile Endzustände: Entweder alle sind am Ende Freunde oder es bilden sich zwei verschiedene Gruppen heraus, die unter sich befreundet sind und mit der anderen Gruppe verfeindet sind.[139, s. 124] In der Praxis ist diese Theorie anhand der Konstellation der Bündnisse vor dem Ersten Weltkrieg untersucht worden. Hier wurde festgestellt, dass als die Bündnisse ein ausgeglichenes Verhältnis erreichten, der erste Weltkrieg ausbrach. Die Bündnisse wurden so lange umgeknüpft, bis ein stabiler Zustand erreicht wurde. Dabei standen sich am Ende zwei untereinander verbündete und gegenseitig verfeindete Bündnisse gegenüber. [139, s. 127]

Vertrauen

Vertrauensnetzwerke sind Freundschaftsnetzwerken ähnlich. Ein Unterschied ist, dass Freundschaften auf gegenseitigkeit beruhen. Vertrauen ist dagegen eine gerichtete Größe. Es ist kein Widerspruch, dass eine Person einer anderen Person vertraut, diese andere Person das Vertrauen jedoch nicht erwiedert. Freundschaften dagegen müssen auf Gegenseitigkeit beruhen.
Dabei spielt auch eine wichtige Rolle, wie Vertrauen an Dritte weiter gereicht wird. Wenn eine Person seinem Nachbarn einen Handwerker empfiehlt, dann gibt diese Person Vertrauen weiter. Mein Vertrauen wirkt sich auf das Vertrauen anderer aus. Dieser Effekt wird jedoch über viele Bekannte schwächer.
In sozialen Netzen wie beispielsweise Facebook kann Vertrauen auch über Zugriffsrechte verteilt werden. In den Privatsphäre-Einstellungen finden sich Freigabeformen wie: Freunde, Freunde von Freunden, weltweite Freigabe. Durch diese einfachen Zugriffsrechte kann die Vererbung von Berechtigungen eingestellt werden. Die Verbindung zwischen mir und meinen Freunden wird zu einer Beziehung, die aus einer Freundschafts- auch eine Vertrauensverbindung schließt. Wenn Nachrichten nur mit Freunden geteilt werden, dann werden Freunde dieser Freunde sie nicht sehen. Außer, die Freunde von Freunden sind auch die eigenen Freunde. Wenn Freunde von Freunden meine Nachrichten sehen dürfen, dann wird das Vertrauen über Freunde an deren Freunde weiter gegeben.
Ein anderes Beispiel für solche Netze findet sich auch im Web of Trust Layer des ursprünglichen Semantic Web Stacks. Hier hat das Web of Trust die Aufgabe, Daten- und Wissenquellen zu bewerten und eine Einschätzung des Inhalts darzulegen. Dies sollte den

Daten im Semantic Web eine qualitative Ebene geben. Gerade im Falle von Daten für die Datenverarbeitung ist es wichtig, vertrauenswürdige Quellen zu verwenden, da schlechte oder fehlerhafte Daten Ergebnisse verfälschen.[5] Die transitive Argumentation in Netzen trifft im Falle des Semantic Web auf zwei Ebenen zu. Im Semantic Web werden idealerweise Daten und Regeln verwendet, um neues Wissen zu generieren. Dies ist eine Fehlerquelle, aus der sich Fehler auch transitiv weiter verbreiten können.
Diese Verfahren setzen gezielt darauf, dass es keine zentrale Vertrauensinstanz gibt. Das Wissen und Vertrauen soll aus der Masse kommen und weitergegeben werden. Die Bewertung einzelner Quellen wird dabei der Schwarmintelligenz oder besser Netzintelligenz übertragen.
Vertrauen kann in diesem Rahmen auch weiter ausgelegt werden. Facebook-Likes[6] sind auch eine Weitergabe von Vertrauen auf gute Beiträge oder guten Geschmack. Das Vertrauen und die Likes werden zum digitalen Kapital.
Im akademischen Bereich können Zitate analog verstanden werden. Je öfter die Werke eines Wissenschaftlers zitiert werden, umso höher ist sein Impact Factor. Ein weiteres Maß, das auf dem Impact Faktor aufbaut, ist der Hirschindex[130]. Im Science Fiction Roman „Down and Out in the Magic Kingdom“[75] wird dieses Prinzip auf die Spitze getrieben und das Vertrauenskapital bestimmt über den Zugang zu Ressourcen und Positionen. Vertrauen ersetzt Geld als Zugangsbeschränkung zu Ressourcen.

Kontrolle, Macht, Herrschaft

Kontrolle im generellen Sinne kann als ad-hoc Zugriff auf das Handeln von jemandem oder etwas anderem gesehen werden.
Oft bilden sich bei Kontrollstrukturen einfache Gebilde. Diese könnten beispielsweise in Organigrammen ausgedrückt werden. Kontrolle hat auch etwas mit Hierarchien zu tun. In klaren Fällen stimmt dies auch. Der Schluss liegt nahe, da man an eine Pyramidenstruktur denkt. Dabei lässt sich Kontrolle einfach von oben nach unten weiterleiten, wie in Abbildung 2.2 dargestellt.
Es entstehen Befehlsketten. A kontrolliert B und B kontrolliert C. A hat indirekte Kontrolle, Befehlsgewalt über C. Dieses Prinzip kann dabei für einige Verwirrung sorgen, etwa dann, wenn aus dem oberen Beispiel C Kontrolle über A hat. Diese Überlegung stört das hierarchische Kontrollsystem.
Beispielhaft kann dies an einer erdachten hierarchischen Struktur einer Firma betrachten. An der Spitze steht der Chef, unter ihm die Manager, unter den Managern die Abteilungsleiter und am Ende stehen die einfachen Angestellten, wie in Abbildung 2.2 dargestellt.
In dieser klaren Befehlsstruktur lassen sich Pfade vom Chef zum einfachen Angestellten ausmachen. Dabei kommuniziert der Chef nicht direkt mit dem Angestellten, sondern steuert diesen indirekt durch seine Manager und diese über die Abteilungsleiter. Diese Struk-

[5] Dies wird in Kapitel 4 besprochen.

[6] Weitere Aspekte von sozialen Netzwerken werden in Abschnitt 2.2.3 diskutiert.

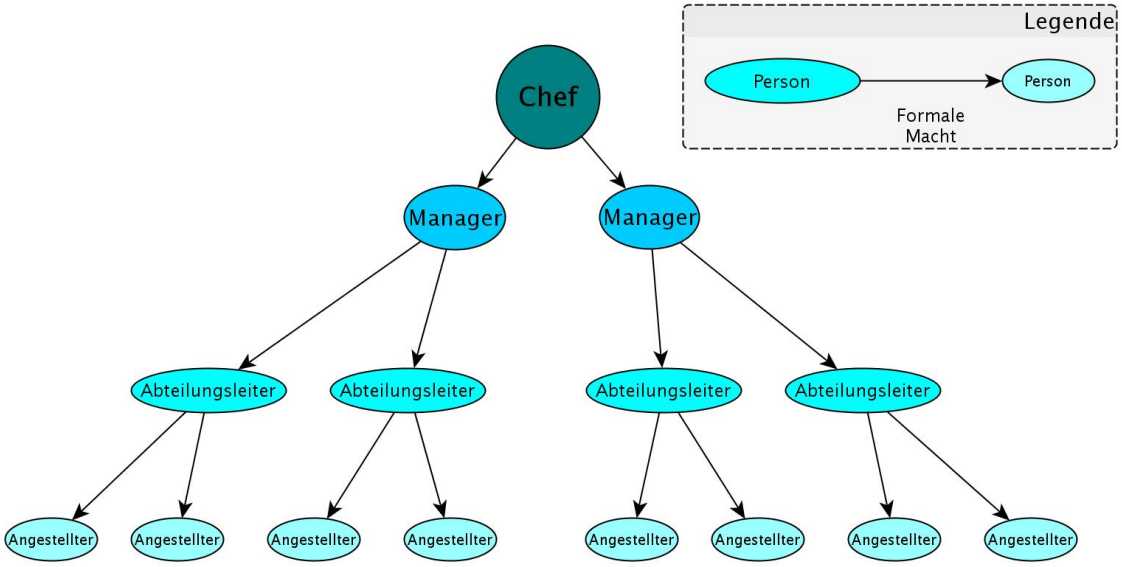


Abbildung 2.1: *Modell einer Firma. Die Befehlsstruktur ist klar hierarchisch organisiert.*

turen können auch durch die Analyse von Kommunikationsprozessen ermittelt werden.[97]

Außerhalb des Modells ist eine so klare und widerspruchsfreie Machtstruktur eher eine Ausnahme. Der Chef kann natürlich direkt Angestellte ansprechen, soweit er diese Möglichkeit nutzt. Arbeitsplätze sind jedoch soziale Räume, in denen nicht nur befohlen und gearbeitet wird. Auch hier gibt es weitergehende soziale Beziehungen. Wenn beispielsweise die Frau des Chefs in der Firma des Chefs angestellt ist, wird diese eine andere Position annehmen als ein einfacher Angestellter, auch wenn dies ihre Position in der Hierarchie wäre.

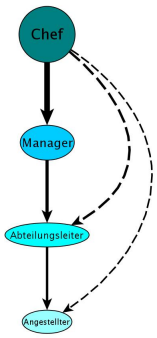


Abbildung 2.2: *Ein Befehls- oder Einflusspfad in einer klaren Hierarchie*

Es kann auch zu anderen, informellen Beziehungen kommen, die die Hierarchien durchbrechen. Wenn beispielsweise ein Abteilungsleiter mit einem einfachen Angestellten regelmäßig Fußball schaut, wird durch diese Freundschaft wahrscheinlich die Art der Machtausübung verändert, dies ist in Abbildung 2.3 Abgebildet.
Dabei wird die Macht nicht aufgehoben. Gewisse Formen der Machtausübung können die Freundschaft gefährden. Die Machtausübung wird sich also verändern. Aber auch generell hat Macht Grenzen und nicht jede Machtausübung kann von den Untergebenen hingenommen werden, ohne dass es Konsequenzen hat.
Max Weber spricht in seiner Definition von Macht und Herrschaft von der Chance eine Handlung auch gegen den Willen einer Person zu veranlassen.[263, Kapitel 1 §16] Herrschaft und Kontrolle müssen nicht als absolut gesehen werden. Tradierte und legale Macht kann dabei von anderen Formen der informellen Macht außer Kraft gesetzt werden. Eine andere Lesart von Chance ist auch, dass die kontrollierende Instanz nicht von ihrer Macht Gebrauch macht. Es könnte von Machtpotenzial, mit einer betonung auf Potential, gesprochen werden.

Ob Macht und Herrschaft über viele Akteure hinweg immer durchzusetzen ist, sollte dabei auch nicht als gegeben angenommen werden. Eine Person hat durch eine Rolle (Abteilungsleiter) Macht, wendet diese aufgrund einer weiteren oder komplexen Beziehung nicht an. Dies wird nach Merton auch als Rollenkonflikt zwischen den Rollen, die eine Person einnimmt, gesehen.[168] Es wird also aufgrund von Rollenkonflikten (Freund und Abteilungsleiter) versucht, einer Person keine Handlungen gegen ihren Willen zu befehlen. Die Dokumentation und das Erfassen solcher Relationen tritt dabei auf zwei verschiedenen Ebenen auf. Die Firmenstruktur kann zentral erfasst sein und von der Firma selbst oder auf der Homepage repräsentiert werden. Die informellen Beziehungen der Angestellten untereinander sind dort eher selten erfasst. Hier könnten Beiträge aus sozialen Netzwerken, das Auftauchen der Angestellten in anderen Kontexten, wie der Website eines Sportvereins die Beziehungen abbilden. Auf einer nicht menschlichen Ebene kann Kontrolle auch

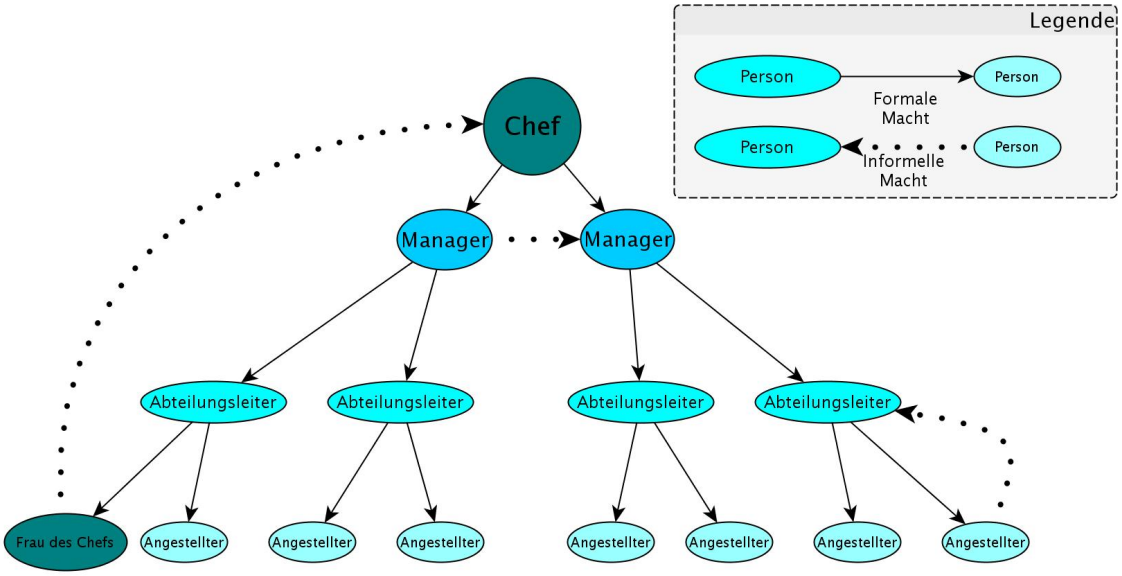


Abbildung 2.3: *Machtpyramide unter Berücksichtigung spezieller informeller Machtverhältnisse. Diese können hier den sonst simplen Fluss der Macht durch Kreise etc. durcheinanderbringen.*

anders definiert werden. Zwischen Firmen bestehen auch Machtverhältnisse. Eine Firma kann einer anderen Firma gehören. Dies ist in verschiedenen Formen möglich. Modellhaft kann von einer Vollkontrolle einer Firma gesprochen werden, wenn eine Akteur mehr als 50 Prozent der Anteile hält. Dann übt dieser Akteur volle Kontrolle über die andere Firma aus.
In einer Studie zu den Besitzverhältnissen zwischen großen Firmen gab es einige wenige Akteure, die direkt und indirekt einen Großteil des weltweiten Kapitals kontrollieren[7]. Dabei ist mit Kapital der Wert eines internationalen Konzerns gemeint.
Hier entsteht ein Fluss aus Einfluss, dessen Quelle oder auch Quellen nicht direkt vom ausführenden Akteur aus zu sehen sind.[256] Interessant für die Datenerfassung ist, dass

[7]Laut Studie sind 737 der Akteure, indirekt im Besitz von 80 Prozent des Kapitals, das in internationalen Konzernen liegt.[256]

diese Informationen aus offiziellen Dokumenten, die veröffentlich werden müssen, hervorgehen und daher kein explizites Geheimwissen darstellen. Auch stellt sich hier nicht die Frage, welche Regeln gelten, da sich Konzerne im Rahmen des Rechts bewegen müssen und daher die Regeln der Kontrolle vorliegen. Interessant ist hier eher Inwiefern diese Kontrolle und der Besitz strategisch genutzt wird.

Teilen und Zerschlagen

Menschliche Beziehungen bilden größere Konstrukte. Menschen finden sich in Organisationen zusammen. Nicht immer ist eine solche Organisation zentral bestimmt wie die hierarchische Struktur im vorausgehenden Beispiel. Solche dezentralen Strukturen lassen sich nicht einfach erfassen, da es keine zentrale Instanz gibt, die ihre Untergebenen dokumentiert. Die Erhebung der Struktur über die Verbindungen ist dabei eine wichtige Komponente.
Gerade terroristische und kriminelle Organisationen haben eine eher lose Gruppenstruktur. Sie haben keine zentral bestimmende Struktur, sondern sind in kleineren Zellen organisiert und hängen nur über wenige Akteure zusammen. Dies hat den Vorteil, dass Teile der Organisation zerstört oder verhaftet werden können, ohne die gesamte Gruppe zu gefährden. Es ist immer nur eine Teilgruppe betroffen. Die Organisation an sich lässt sich auf diese Weise nicht komplett auflösen. [73]
Ein Beispiel für eine solche organisatorische Struktur sind kriminelle Motorradklubs. Es gibt keine wirklich rechtliche Handhabe gegen alle Chapter eines kriminellen Motorrad Klubs, da dies rechtlich eigenständige Einheiten sind. Ein Verbot kann so nur einzelne lokale Chapter treffen und nicht die gesamte Organisation.[195] Ein Netzwerk ist es trotzdem, da die kleinsten Einheiten nicht vollkommen unabhängig voneinander agieren. Die einzelnen Chapters gehören immer noch zu einer Gesamtstruktur.
Exekutive und Judikative versuchen, solche kriminellen Netzwerke zu zerschlagen. Dies soll effektiv durch das Herausnehmen von Cutting-Points, zentralen Figuren im Organisationsgeflecht geschehen.[164] Da Verbrecher und Terroristen meist keine Fragebögen ausfüllen, müssen diese Verbindungen zwischen den Akteuren in solchen Netzen erst einmal recherchiert werden.
Dabei können große Datenmengen anfallen beispielsweise beim Aufzeichnen von Kommunikationsmetadaten, Treffen und Übereinstimmungen in Adressbüchern.

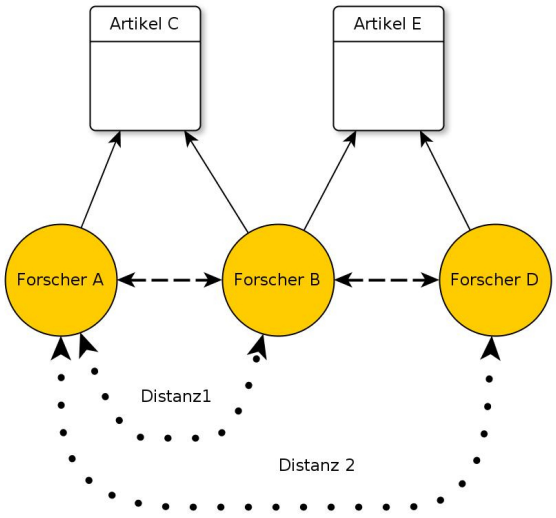
2.2.2 Der Mensch verstrickt mit allem Anderen

Während die Sozialwissenschaft auf ein festes Instrumentarium an Befragungsmethoden und Datenerhebungen zurückgreifen kann, bieten auch andere Herangehensweisen die Möglichkeit, Netze zu konstruieren. Die Schlussfolgerung eines Zusammenhangs und eines sich daraus erstellenden Netzwerks bezieht sich auf die Beobachtung von dritten Sachverhalten, die dazwischen liegen und sichtbar sind, ohne dass die untersuchten Beziehungen durch direkte Befragung zustande kamen.

Kollaboration

Im Abschnitt 2.2.1 zum Small-World Phänomen wurde das Phänomen einer kleinen Welt von Bekanntschaften betrachtet. Diese Small-World Theorie lässt sich auf anderen Datensätzen recht einfach beobachten. Hier wird nicht die ganze Menschheit genommen, sondern nur ein Ausschnitt, eine spezialisierte Community.
Der ungarische Mathematiker Paul Erdős hat beispielsweise den Begriff der Erdős-Zahl geprägt. Diese beruht auf der kollaborativen Distanz zu Paul Erdős. Kollaborative Distanz kann anhand der gemeinsamen Publikationen beschrieben und gemessen werden. Das verbindende Element ist dabei die direkte Zusammenarbeit an einem wissenschaftlichen Artikel.[20] Forscher A hat mit Forscher B an Artikel 1 gearbeitet. Forscher A und

Abbildung 2.4: *Beispiel für die Errechnung von kollaborativer Distanz*



Forscher B haben damit eine Kollaborativdistanz von 1.
Forscher A und Forscher B haben zusammen an Artikel 1 gearbeitet und Forscher B und Forscher C haben zusammen an Artikel 2 gearbeitet. Forscher A und Forscher D haben eine Kollaborativdistanz von 2, unter der Bedingung, dass Forscher A und Forscher D nicht Artikel 3 zusammen veröffentlicht haben. Der Sachverhalt ist in Abbildung 2.4 abgebildet. Die Erdős-Zahl besagt, dass das eine Ende der Kette immer Paul Erdős ist. Die Person, der eine Erdős-Zahl zugeordnet wird, ist das andere Ende der Kette. Hier wird die Länge des kürzesten Pfades durch einen Graphen beschrieben. Publikationsnetzwerke, die so entstanden sind, lassen sich viel besser beobachten als die Netzwerke zwischen allen amerikanischen Einwohnern. Sie sind besser verfolgbar, da die Informationen offen zugänglich sind und zur erfassung der Informationen nicht der Schutz der Privatsphäre beachtet werden muss.
Ein Beispiel aus der Populärkultur ist das Oracle of Kevin Bacon. Diese Website wurde auf Grundlage der Internet Movie Database (IMDB)[135] erstellt. Die Internet Movie Database ist eine Datenbank über Filme. Neben Bewertungen von Filmen, Erscheinungsdaten und anderen Informationen enthält diese Datenbank auch die Zuordnung von Schauspielern, Regisseuren und anderen Beteiligten an Filmen. Als Nachschlagewerk konzipiert besteht

in der Zuordnung von Schauspielern zu Filmen eine nicht beachtete Informationsquelle. Die allgemeine Theorie der Small-World[8] kann mit diesen Daten veranschaulicht werden. Wie bei der Betrachtung der mathematischen Publikationen wurden die Filme, bei denen eine Kollaboration stattfand, zu den Kanten zwischen den Personen hier den Schauspielern.
Eine Kombination aus zwei Knoten und einer Kante wurde damit zur Aussage: Schauspieler A spielte zusammen mit Schauspieler B in Film C. Hier nehmen die Schauspieler die Knoten ein und der Film C ist dabei die Kante.[9]
Kevin Bacon stellt einen willkürlichen Mittelpunkt des Netzes dar, genauso wie Paul Erdős. Die perfekten Zentren von Netzwerken lassen sich nur zum Zeitpunkt der Betrachtungen feststellen, da nach wie vor mathematische Publikationen und Filme erscheinen. Eine Kombination aus beiden Nummern zeigt eine andere Art einer "Bewertung" des Menschen anhand seines Umfelds. Dies ist dann die Erdős-Bacon-Zahl. Diese zeigt durch ihr reines Bestehen an, dass eine Person Teil zweier Welten ist, der mathematischen und der des Films. Hier wird die niedrigste Erdős-Bacon-Zahl u.a. von Bruce Reznick belegt. Er hat mit Erdős publiziert und weiterhin eine Bacon-Zahl von 2. Daraus ergibt sich insgesamt eine Erdős-Bacon-Zahl von 3. Erdős selbst hat nur eine Erdős-Bacon-Zahl von 4. Er taucht in einer Dokumentation über sein Leben auf und hat daher die Bacon-Zahl von 4 und selbst die Erdős-Zahl 0.[232]
Bei den Zusammenhängen über Kollaborationen wurden auch die akkumulierten Daten aus dem Patentwesen analysiert. Dabei wurden aus den Patenten Verbindungen zwischen Ländern errechnet. Die einzelnen Personen verlieren in diesem Zusammenhang ihre Bedeutung.[114]
Auf diese Weise können auch Korrespondenzen untersucht werden. Als Netz können diese Korrespondenzen Aufschluss über den Austausch von Ideen geben. Visualisierte Netze bieten ein Mittel um zugang zu den konkreten Dokumenten zu schaffen.[53]

[8] Siehe Abschnitt 2.2.1.

[9] Eine Kante kann dabei mehr als einen Film repräsentieren, denn Schauspieler A kann mit Schauspieler B in einer ganzen Reihe von Filmen mitgespielt haben.

Der Mensch und die Dinge

Neben der Kollaboration von Menschen über Werke können Verbindungen jeglichen anderen Zusammenhangs auf diese Weise hergestellt werden. Ein weiteres Beispiel sind Dinge. Hier ist der Handel hervor zu heben. Menschen kaufen Dinge. Diese Kaufentscheidungen werden aufgezeichnet. Dabei entstehen Netze von Menschen, den Konsumenten und Dingen, den Produkten. Diese Daten sind einfach zu erheben. Sie entstehen als informationstechnisches Nebenprodukt eines Bestellprozesses oder eines Kaufs mit eindeutig zuordnenbarer Person. Die Bestellung eines Gegenstands erzeugt eine einfache Assoziation zwischen dem Besteller und dem bestellten Produkt. Hieraus können die Ähnlichkeiten zwischen Produkten und den Ähnlichkeiten zwischen Konsumenten ermittelt werden. Diese Ähnlichkeiten bilden Netze. Sie können dazu genutzt werden, dem Nutzer weitere Artikel vorzuschlagen. Es handelt sich also um ein Empfehlungssystem[10], dass auf der Grundlage von in der Vergangenheit gleichförmig getätigten Kaufentscheidungen Empfehlungen anzeigt.[274, 276]

2.2.3 Das WWW als Netz auf dem Netz

Netze ganz anderer Art bilden Übertragungsnetze wie das Internet. Doch bildet das weltweite Übertragungsnetz mit seinem technischen Hintergrund die Grundlage für ganz andere Arten von Netzen wie dem World Wide Web kurz WWW, einem Netz aus Hypertext-Dokumenten.
Das Menschenlesbare WWW basiert daher auf mehreren Schichten von Netzen. Es gibt zum einen die physische Infrastruktur aus verschiedenen Technologien wie Lichtwellenleitern, Kupferleitungen und Netzwerk-Hubs, zum anderen die Adress- oder IP-Ebene, die logische Netze als Abstraktion auf die darunter liegende Hardware abbildet. Auf der abstrakteren Ebene ist das Internet ein Informationsnetz, das in seiner Struktur von der physischen Infrastruktur abstrahiert wird. Auf der Adressebene muss die physische Infrastruktur nicht bekannt sein. In diesem Netz aus Adressen können beliebige Informationen ausgetauscht werden. Das darauf aufbauende World Wide Web ist nur ein Dienst von vielen, wenn auch ein sehr populärer und ein sehr präsenter.
Hyperlinks zwischen Websites bilden ein Netz aus Hypertexten. Diese können als ein gerichtetes Netz angesehen werden. Anders als der physische Unterbau, bei dem eine Leitung meist nicht nur in eine Richtung senden kann, sondern Nachrichten in beide Richtungen versandt werden können, ist ein Hypertext-Link immer eine gerichtete Verbindung auf eine andere Seite oder Ressource. Auf der anderen Seite des Links angekommen führt nicht zwangsläufig ein Weg zurück.
Hypertext, wie der Nutzer ihn in der Form von HTML sieht, ist auch eine Ansammlung von Links. Im Hypertext können beispielsweise Bilder von einem anderen Server verlinkt werden. Eine Website kann an vielen verschiedenen Orten auf der Welt liegen und erst die zusammengefassten Informationsströme bilden die menschenlesbare Website. Für den Be-

[10] Engl.: recommender system

trachter der Webseite ist dies in der menschenlesbaren Representation durch den Browser nicht ersichtlich.

Das Übertragungsnetz

Das Internet ist keine beliebige Netzstruktur, in der sich Knoten beliebig verknüpfen können. Es bildet sich aus einer Menge von kleineren Netzen. Es gibt Local Area Networks (kurz LAN), Metropolitan Area Networks (kurz MAN) und Internet Backbones. Dabei stellt das LAN, das lokalste Netz dar, während das Backbone-Netz ist das geografisch umfassendste Netzwerk ist.[12, s.7 ff]
LANs werden von Unternehmen, Institutionen und Universitäten genutzt, aber auch Netze in Privathaushalten sind LANs. Meist wird der Zugang zu solchen Netzen von Firewalls vor unberechtigtem Zugang geschützt. Auch Verbindungen zu bestimmten Zielen können für die Computer im lokalen Netz gesperrt werden.
Auf der logischen Seite wird anders gearbeitet. Hier werden die Informationen je nach Adresse und Netzadresse versandt. Vereinfacht kann gesagt werden, dass ein Rechner weiß, in welchem Netz er sich befindet. Wenn ein Paket nicht ins eigene Netz gesendet wird, dann muss es an das Gateway des Netzwerkes geschickt werden. Das Gateway kümmert sich um die Weiterleitung ins richtige Zielnetz.[12, s. 66 ff.] Dieses Verfahren wird Routing genannt. Die Informationen werden beim Transfer in Pakete verpackt, die nacheinander über das Netz versandt werden. Daraus entsteht ein Fluss an Informationspaketen, die durch die richtige Weiterleitung den Weg durch das Netz finden. Das hat den Vorteil, dass es keine zentral steuernde Instanz geben muss, die den Informationsfluss koordiniert.[11] Diese Informationspakete suchen sich ihren Weg durch das Netz, ähnlich dem Verfahren im Small-World-Experiment.[12]
Das Internet ist als militärisches Netzwerk geplant worden und daher sehr robust und unanfällig gegen das Ausfallen von Knotenpunkten. Es müsste ein Großteil der Knoten zerstört werden, um das Kommunikationsnetz im Ganzen zusammenbrechen zu lassen.[55] Im Vergleich mit Abschnitt 2.2.1 ist das Internet so konzipiert, dass es eigentlich keine Cutting-Points gibt, an denen es auseinanderbrechen könnte. Das Internet ist auch auf seine Netzstruktur untersucht worden. Dabei wird ein Problem bei der Untersuchung von dezentral strukturierten Netzen sichtbar: Es gibt keine Teilnehmerliste. Um eine Einschätzung der Struktur zu gewinnen oder auch eine Netzkarte des Internets anzulegen, muss gemessen werden. Das Messen einer Route durch das Netz erfolgt dabei u.a durch das Nachvollziehen der Paketrouten. Dabei werden jedoch nur die Routen erfasst, die zwischen den untersuchten Punkten liegen. Das Messen wird meistens durch das Programm *traceroute* realisiert. Das Programm *traceroute* zeigt den Weg eines Pakets durch das Internet an. Dabei muss dieser Weg nicht immer dem kürzesten Pfad entsprechen, was unter anderem am allgemeinen Verkehrsaufkommen im Netz liegen kann.[66] Die Güte des Ergebnisses der Kartografierung hängt dabei auch davon ab, wie viele und welche

[11] Der Weg, den ein Paket nimmt, um von A nach B zu kommen, lässt sich mit Standardsoftware wie "traceroute" verfolgen.

[12] Siehe Abschnitt 2.2.1.

Ausgangs- und Zielpunkte für die Routen gewählt wurden.[66, 6] Vollständige Ergebnisse können mit dieser Herangehensweise nicht erzielt werden. Dabei ist eine Definition des Internets und der Routen auch von der Einbeziehung der Client Server Architektur des Internets abhängig.

Client Server Architektur

Im Internet ist nicht jeder Computer von sich aus gleichberechtigt. Die meisten Menschen, die im Internet unterwegs sind, bilden dabei keine Datenknoten. Clients beantworten im Generellen keine Anfragen von außen, sondern fragen an.
Im Allgemeinen kann von Client-Systemen und Server-Systemen gesprochen werden. Ein Server stellt Dienste wie Websites und andere Kommunikations- und Informationsdienste bereit, während der Client Dienste in Anspruch nimmt, ohne selbst Dienste bereitzustellen.
Im Generellen haben Server einen menschenlesbaren Domänennamen[13]. Des Weiteren zeichnen sich Server dadurch aus, dass sie feste Adressen haben. Sie sind immer gleich zu identifizieren. Clients bei großen Internetprovidern bekommen meist je nach Verbindung eine Adresse aus einem Adresspool. Clients sind somit nicht immer gleich adressierbar und identifizierbar.
Server könnten dabei auch als Austausch- und Kommunikationsmittelpunkt gesehen werden. Die Daten gehen von den Clients an den Server und dieser verteilt sie dann an weitere Clients wie beispielsweise bei einem Gästebuch. Der hier aufgestellte Vergleich basiert auf der Unterscheidung in Intention und Nutzen eines Gerätes. Prinzipiell sind auch ein Client-Rechner dazu in der Lage, Serverprogramme auszuführen. Die strikte Trennung kommt aus einer Zeit, in der die Rechen- und Speicherressourcen ungleich verteilt waren.
Ein Fall, in dem das Client-Server-Pattern aufgehoben ist, sind Peer-to-Peer Netze. Hier vermittelt der Server nur zwischen den Clients. Durch seine feste Adresse ist er nur die Vermittlungsstelle für den Kontakt. Nach der Vermittlung übernehmen die Clients selbst den Datenaustausch.

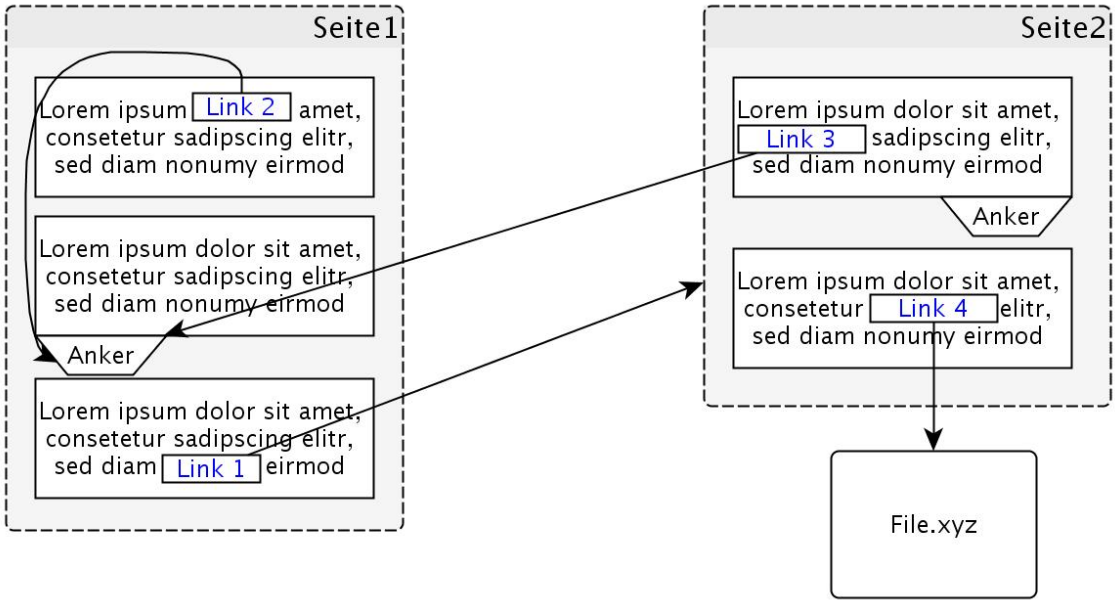
Der Aufbau des Hypertexts

Das World Wide Web ist eine der massentauglichsten Anwendungen des Internets. Durch menschenlesbare Inhalte wurde die Popularität des Internets maßgeblich gesteigert. Dazu trug unter anderem der HTML Standard als Austauschformat für Text und Hypertext bei. Entwickelt wurde dieser Dateistandard von Tim Berners-Lee, welcher auch das Konzept der URL[156][14] entwickelte.
Dies ist die technologische Grundlage, die die Vernetzung dieser Informationsressourcen in HTML erst möglich machte. Aus einer Menge von Hypertext, der auf Webservern angelegt wird, ist ein riesiges Informationsnetz gewachsen.

[13] Die Auflösung von Namen zu Nummern geschieht über den Domain Name Service.

[14] Uniform Resource Locator [156]

Abbildung 2.5: *Hier wird gezeigt, welche Möglichkeiten es im HTML Standard gibt, zwischen HTML Dokumenten zu verweisen.*



In diesem Informationsnetz gibt es Assoziationen durch Links zu diversen Inhalten. Bei der Bewertung und Erschließung kommen Netzwerkanalyseverfahren zum Einsatz.

Hyperlink

Hyperlinks im HTML Standard können auf verschiedenste Arten von Ressourcen verweisen, dabei auch auf sich selbst oder Teile von sich selbst wie in Abbildung 2.5 dargestellt ist. Es gibt mehrere Möglichkeiten der Verknüpfung:

1. ein anderes Hypertext Dokument;
2. eine Stelle im eigenen Dokument, ein Anker;
3. eine Stelle in einem anderen Hypertext Dokument über einen Anker;
4. eine beliebige andere Ressource, meist eine Datei oder Ähnliches, an dieser Stelle wird der Kontext des Hypertexts verlassen;

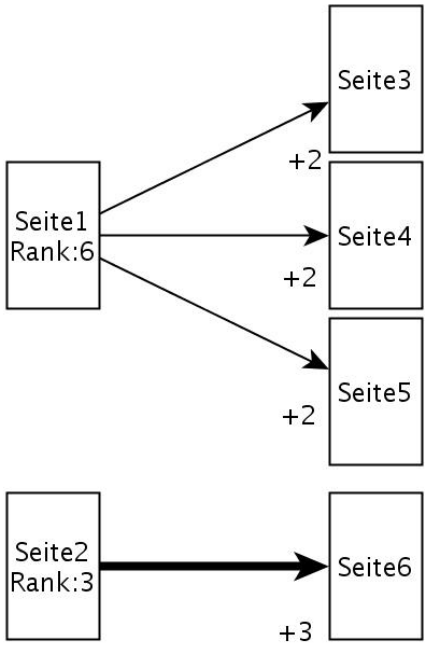
Im HTML-Standard wird ein einfaches “a” Tag mit dem Parameter “href” verwendet, um einen Link im Fließtext auszuzeichnen. Dieses Tag enthält nicht zwangsweise eine Information über die Intention eines Links oder dessen Bedeutung. [196]
In seiner Verwendung als Teil eines Dokumentes wird die Bedeutung des Links durch den Kontext seiner Position beschrieben. Die Semantik des Links ist dem menschlichen Leser des Dokumentes wahrscheinlich ersichtlich, technisch wird die Bedeutung jedoch nicht explizit hinterlegt.
Ein Ansatz, in HTML-Dokumenten Links mit Sinn zu anzureichern und damit HTML maschinenlesbarer zu machen, gab es mit RDFa. Dies ist eine Erweiterung von HTML um Typisierungen der Links. RDFa basiert auf RDF und kann zum Annotieren von Webseiten

in Maschinen lesbarer Form genutzt werden.[4]
RDFa schlägt einen Übergang vom klassischen Netz für die Menschen zu Linked Data[15].

Bewertung des WWW

Der Page-Rank Algorithmus[193] von Larry Page und Sergei Brin beschreibt eine Bewertung von verlinkten Quellen und somit deren Priorisierung in der Suchergebnisanzeige.[16]
Das Prinzip beruht darauf, dass eine Seite ihren eigenen guten Namen durch ihre Links

Abbildung 2.6: *Verteilung von Page-Rank Bewertungen zwischen Websites. Bei vielen ausgehenden Links bekommt jede Seite nur wenig des Page-Ranks ab.(Seite1 zu Seite3,Seite4,Seite5) Sind wenige Seiten verlinkt bekommt jede verlinkte Seite mehr Bewertung ab.(Seite2 zu Seite6)*



weitergibt. Das Vertrauen oder die Relevanz die eine Seite besitzt wird durch Links weiter gegeben. Ein Link von einer Website mit hohem Page-Rank ist damit mehr "wert" als der Link einer Seite, die selbst nur einen niedrigen Page-Rank besitzt, also entweder neu ist oder nicht von anderen Websites verlinkt wird.
Des Weiteren ist die Weitergabe des Page-Ranks durch Links an die Anzahl der Links gebunden, die von der besagten Seite ausgehen. Wenn eine Seite eine große Menge von Links hat, dann wird der Page-Rank von der einen Seite an alle verlinkten Seiten verteilt und jede verlinkte Seite erhält nur wenig "Page-Rank". Links von Seiten, die wenige Links haben, werden gewichtiger. Ihr Page-Rank-Wert verteilt sich auf nur wenige Seiten. Das ist in Abbildung 2.6 dargestellt.[193]
Der hier beschriebene Algorithmus wurde über die Jahre verfeinert und wird nicht mehr

[15]Linked Data wird im Kapitel 4 genauer vorgestellt.

[16]Page-Rank wurde auch lange als Hauptgewichtungsverfahren bei der Firma Google benutzt, welche die Erfinder des Algorithmus gründeten.

in der Ursprünglichen form verwendet. Das wurde unter anderem durch den Missbrauch des Verfahrens notwendig. Beispielsweise wurden automatisch erstellte Links in Kommentaren dafür genutzt, die eigene Website aufzuwerten.[203][17]. Der Algorithmus basiert auf der Annahme von relativ statischen Dokumenten und nicht von dynamischen, interaktiven Websites mit nutzerbasierten Daten. Dass der darunterliegende Algorithmus bekannt ist, verleitet natürlich gerade im Internet, in dem Aufmerksamkeit ein knappes Gut ist, zum Missbrauch des Bewertungsverfahrens.
Page-Rank ist einfach und kostensparend zu realisieren. Deshalb war er anfangs erfolgreich. Jede Website wird einmal besucht, die Links werden analysiert und im nächsten Schritt werden die neuen Page-Ranks berechnet. Neue Seiten werden mit einem fixen Einstiegsranking berücksichtigt. Das Verfahren hat vor allem den Vorteil, dass das Verschwinden von Links, Webseiten etc. den Ablauf und die weitere Berechnung nicht beeinträchtigen, sondern diese ohne Probleme im nächsten Verarbeitungsschritt mit einbezogen werden.
Ähnlich der Datenerhebungen in der Soziologie wird hier versucht, durch indirekte Beobachtung der Links die Wichtigkeit von Information zu erkennen. Dabei gilt aber die Grundprämisse, dass Links immer Empfehlungen sind. Ein Link auf ein Negativbeispiel oder der Link als negatives Element wird bei diesem Verfahren nicht wirklich beachtet. Aber wie bei anderen Beobachtungen ist dieses Verfahren nur gut einsetzbar, solange die Beobachtung und die daraus hervorgehenden Bewertungskriterien nicht bekannt sind und misbraucht werden.

Soziale Netze und WWW

Aus der Analogie des sozialen Netzwerks als theoretische Beschreibung des menschlichen Umfelds sind in den letzten Jahrzehnten explizite informatische Systeme geworden. In diesen sozialen Netzwerken können Menschen miteinander in Kontakt treten. Beispiele aus den letzten Jahren sind dabei Facebook, twitter, MySpace, StudiVz, Friendster etc.
Wie real Personen in diesen Netzen sind und ob nun nur genau ein Mensch eine virtuelle Person sein kann, ist dabei mehr als unsicher. Da Menschen nicht immer allen Menschen in ihrem Umfeld gleich gegenübertreten, führt eine Fixierung auf eine einzige virtuelle Repräsentation zu Problemen.
In sozialen Netzwerken lassen sich direkte Verbindungen, meist Freundschaften genannt, von den Nutzern, meist in gegenseitigem Einverständnis, anlegen. Dabei muss ein virtueller Freund nicht gleichbedeutend mit einem realen Freund sein. Freundschaften sind in den meisten sozialen Netzwerken jedoch der Dreh- und Angelpunkt der Funktion.
In Facebook können Nutzer ihre Posts, Bilder, Adresse und weitere Informationen durch Freundschaftsbeziehungen freigeben. Dabei können Freunde und Freunde von Freunden gewählt werden, mit denen man einigen Content teilt. Des Weiteren sind Einschränkungen und Freigaben möglich. Auch das Verschicken von Nachrichten und die allgemeine Sichtbarkeit des Profils können über diesen Mechanismus gesteuert werden.

[17] Darauf reagierte Google mit dem “nofollow” im HTML Link Tag. Damit wurden Links automatisch von der Bewertung im Page-Rank ausgenommen.[203]

Im Social Network Friendster wurde ein ähnlicher Mechanismus auf andere Weise genutzt. Hier konnten nur Menschen gefunden werden, welche über 4 Ecken bekannt waren, die daher Freunde von Freunden von Freunden von Freunden waren. Diese Einschränkung erinnert dabei an die Small-World Theorie. Durch diese Einschränkungen wurden viele zweckmäßige Freundschaften geschlossen, welche nur zur Erweiterung des Suchradius dienten. Dabei kam es auch zur Erstellung von nicht Personen zugehörigen Profilen, die nur den Zweck hatten, den Degree of Seperation zu senken.[37] Das bisher erfolgreichste soziale Netzwerk ist das 2004 von Marc Zuckerberg gegründete Facebook. Facebook ist seit 2012 ein börsennotiertes Unternehmen und hat mehr als 728 Millionen täglich aktive Nutzer im 3. Quartal 2013[91]. Dies ist ein enormes Datenaufkommen. Die Welt von Facebook ist geschlossen. Es können keine Daten einfach herausgeholt werden. Es ist eine Blackbox, genereller Zugriff von außen ist nicht möglich. Dies entzieht es der öffentlichen, damit der vom Anbieter unkontrollierten Netzwerk-Forschung. Es ist also für den Forscher nicht ohne Weiteres möglich, an die Daten von Facebook zu gelangen.
Im Weiteren werden zwei Verfahren vorgestellt, Netzwerke aus sozialen Netzwerken wie Facebook zu erstellen:
Da wäre als erster Ansatz das Crawling zu nennen. Crawling extrahiert, im Fall von Facebook, Netzwerke aus den Freundeslisten im Web-Frontend. Dabei wird eine Person als Mittelpunkt genommen. Von diesem Mittelpunkt aus werden alle Freunde abgefragt. Dann werden die Freundeslisten dieser Freunde abgefragt und so weiter.[51]
Dieses Vorgehen entspricht im informatischen Sinne einer Breitensuche und zeigt nur die Daten an, die der dazu verwendete Nutzer sehen darf. Hier fallen Nutzer heraus, deren Freundeslisten der Account, mit dem das Crawling durchgeführt wird, nicht sehen darf. Das Crawling ist jedoch offiziell durch die Geschäftsbedingungen von Facebook untersagt, es stellt jedoch methodisch ein verbreitetes exploratives vorgehen der Netzwerkforschung dar.
Andere Verfahren legen mehrere Ego-Netze zusammen[18]. Dabei werden nur Freunde und Freunde von Freunden ausgelesen. Durch die Überschneidungen mehrerer dieser Ego-Netze wird ein Gesamtnetz erzeugt. Um auf diese Weise größere Netze zu extrahieren, müssen mehrere Nutzer ihre Daten zusammenlegen. Dieses Vorgehen ist in der Facebook App [19] des Oxford Internet Institutes implementiert. Dieses Vorgehen ist nach den Geschäftsbedingungen von Facebook legal.
Beim Kombinieren der Ego-Netze entstehen kleinere Netzausschnitte, als sie vom Crawling erzeugt werden. Das Crawling erzeugt einen weitläufigeren Ausschnitt des Gesamtnetzes, dafür sind die Abschnitte aus den kombinierten Ego-Netzwerken vollständiger. Der Vorteil des Verfahrens liegt darin, dass nicht nur von einem Nutzeraccount als Ausgangspunkt aus analysiert wird. Es werden eine Menge von Punkten genommen; diese haben von ihrer Masse aus eine Menge mehr Zugriffsrechte und erzeugen daher vollständigere Daten.

[18]Das Konzept des Ego-Netzwerks wurde in Abschnitt 2.2.1 erläutert.

[19]Quellcode abrufbar unter `https://github.com/oxfordinternetinstitute/NameGenWeb` zuletzt gesehen am 03.07.2014

2.2.4 Straßen, Routen und Geografie

Netze bilden sich auch zwischen Orten, im geografischen Raum. Diese Netze werden u.a. für das Planen von Reisen oder der Organisation von Warenaustausch genutzt. Informatisch wird dabei oft der kürzeste Pfad ermittelt. Konkret können dies verschiedenste Pfade sein: der schnellste, billigste oder zuverlässigste Pfad zwischen zwei Orten. Mit einer Verbindung zwischen zwei Orten muss nicht zwangsweise eine befestigte Straße gemeint sein. Es kann sich dabei auch um einen Flug-Korridor oder eine Schifffahrtsroute handeln.
Eine strikte Simplifizierung auf eine Entfernung von Punkt A nach Punkt B verbietet sich meistens. Selbst Flugzeuge haben Luftsicherheitsmaßnahmen einzuhalten. Bei Booten, Autos und Radfahrern gibt es andere Aspekte zu beachten. Eine Strecke kann nur langsam zu befahren sein, da sie aufgrund des Zustands der Strecke unsicher ist, von rücksichtslosen Verkehrsteilnehmern genutzt wird oder dort Überfälle stattfinden. Auch sind nicht alle Strecken für alle Verkehrsmittel tauglich. Die Einschränkung vom Bootsverkehr bezieht sich auf Gewässer und diese müssen auch je nach der Beschaffenheit des Bootes schiffbar sein. So entwickelt sich der Weg oder die Route weg von einem simplen Überwinden von geografischer Distanz.
Ein Netz aus diesen Informationen bietet die Möglichkeit, algorithmisch einen Weg zu finden, der am kürzesten, billigsten oder schnellsten ist. Um diese Probleme zu lösen, muss ein Netz vorhanden sein, indem hinterlegt ist welche Strecke welche Kosten verusacht, welche zulässige höchstgeschwindigkeit gilt oder welche Risiken vorherrschen etc.

Historische Routen

In der historischen Forschung gibt es Untersuchungen von Handelsnetzen und Reiserouten. Handelsrouten wurden beispielsweise für das Römische Reich untersucht.[136] In dieser Zeit gab es einen sehr ausgeprägten Handel um das Mittelmeer herum. Ein wichtiger Grund dafür war die Versorgung von Rom als Hauptstadt des Reiches mit Nahrungs- und Genussmitteln. Die Stadt Rom hatte solche Ausmaße, dass sie nicht einfach durch ihr Umfeld ernährt werden konnte. Also war sie zur Nahrungsversorgung ihrer Bevölkerung auf Importe aus ihren Provinzen, wie beispielsweise aus Nordafrika angewiesen.
Anders als heutzutage müssen bei der Betrachtung der antiken Welt einige Grundannahmen getroffen werden. So ist eine großflächige rationale Aufnahme von geografischen Distanzen zu dieser Zeit wahrscheinlich nicht möglich gewesen. Exaktes geografisches Wissen in der Form, wie es heute vorliegt, war damals wahrscheinlich noch nicht vorhanden.[20][262]
Die Frage nach der realen Distanz der Schiffbarkeit und Ähnlichem ist schwer zu stellen. Aus dokumentierten Erfahrungen kann jedoch auch ein Navigationsnetz gebaut werden.
Eine Kombination einer Vielzahl von Erfahrungswerten aus historischen Quellen bildet die Datengrundlage des ORBIS-Projekts der Stanford University[218]. In diesem Projekt wurde ein Navigatiossystem für das Römische Reich um das Jahr 200 nach Christus entwickelt. Die Informationen zu den Routen werden dabei aus Erfahrungswerten und historischen Aufzeichnungen gewonnen. Diese Informationen bilden ein komplexes Netz aus

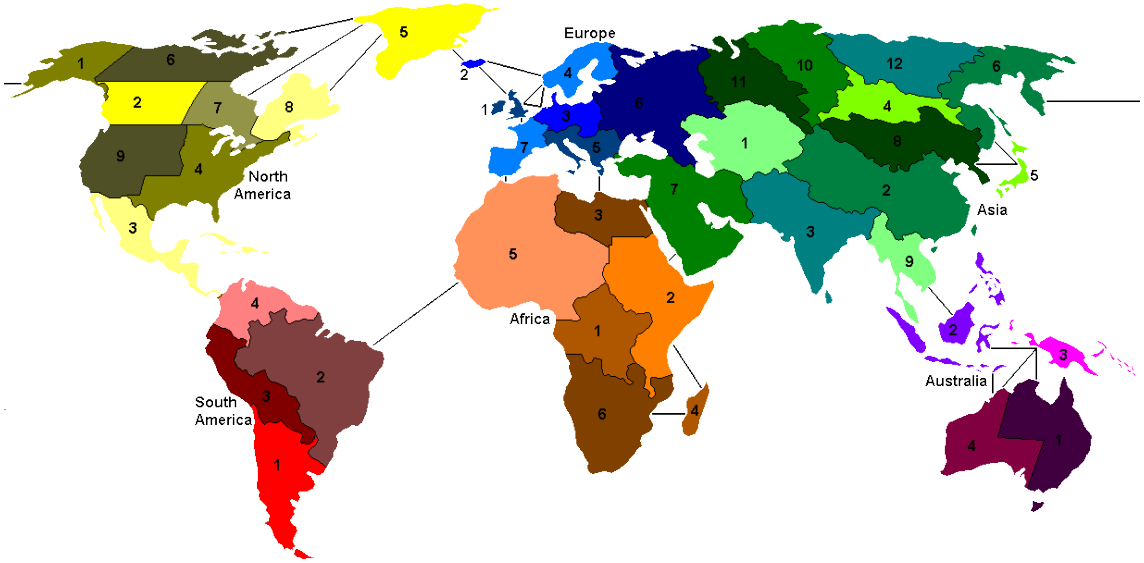
[20] Ein Beispiel dafür ist die “Tabula Peutingeriana”. Sie wird in Abschnitt 2.4.2 besprochen.

möglichen Verbindungen. Die Informationen, die zum Netz beitragen, sind unabhängig voneinander entstanden, lassen sich aber zu einem Netz zusammenführen. Ein wirkliches Distanz- und Routennetz ergibt sich erst, wenn die Rahmenbedingungen angegeben werden. Dabei ist in ORBIS auch berücksichtigt, wie sich die Jahreszeit auf die Passierbarkeit von Routen und auf die Reisegeschwindigkeit auswirkt. Da das Boot eine der wichtigsten Transportmöglichkeiten zu dieser Zeit war, ist die Simulation in diesem Bereich wichtig, da Routen nur bedingt bekannt sind. Es können daher keine klaren Annahmen über die wirklich verwendeten "Wege" getroffen werden, jedenfalls nicht so einfach wie bei Straßen. Deren Verlauf kann anhand von Ausgrabungen bestimmt werden.
Daher müssen die Routen simuliert werden. Dies beinhaltet beispielsweise, dass Schiffe nur in Sichtweite des Ufers fuhren. Des Weiteren ist die Schifffahrt im Winter gefährlicher als im Sommer. Daher sind diese Strecken im Winter nicht so effektiv wie im Sommer.

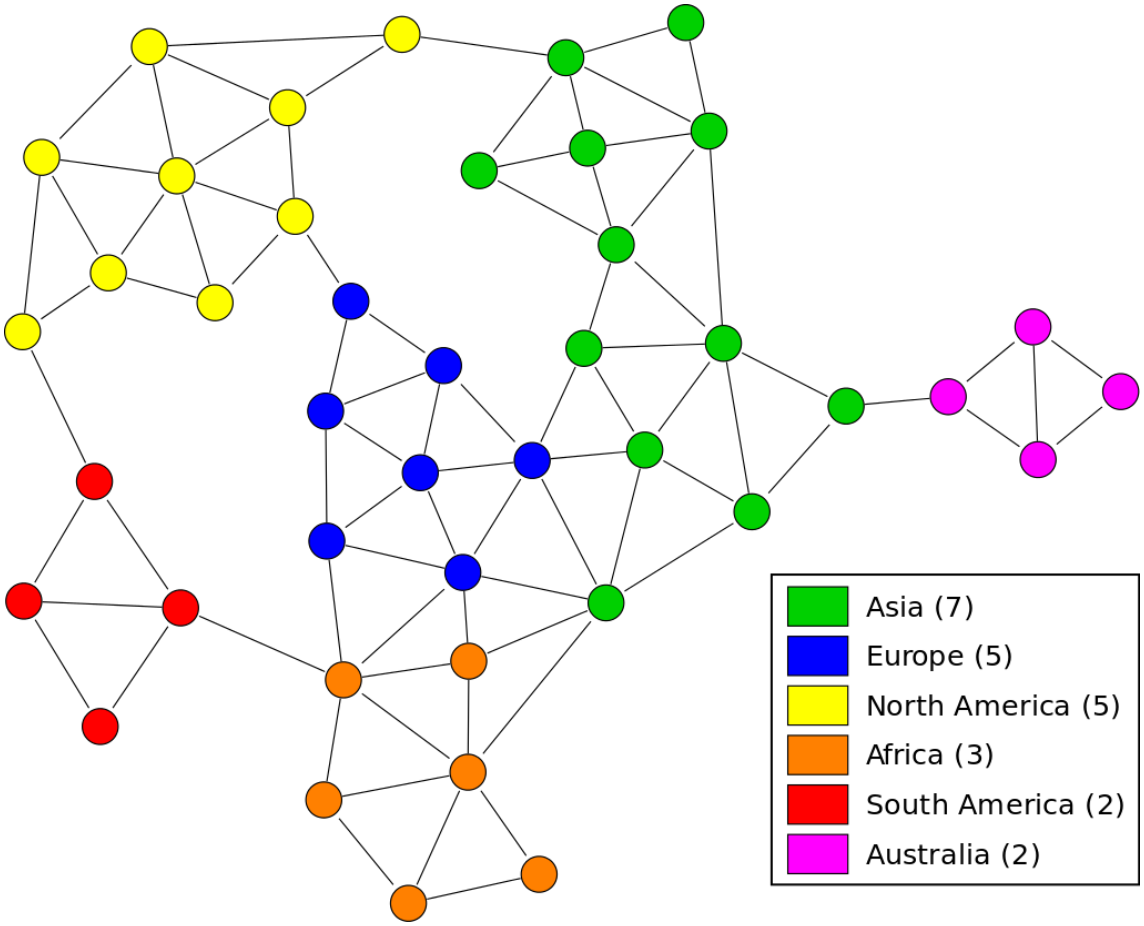
Das Netz als Handlungsfeld im Spiel

Risiko[152] ist ein Strategie-Brettspiel. Der Spielplan von Risiko bildet eine Weltkarte ab, ähnlich Abbildung 2.7a. Das Ziel von Risiko ist es, entweder Missionen zu erfüllen z.B. Kontinente zu erobern oder alle anderen Gegner zu besiegen. Gebiete werden mit Armeen erobert. Armeen können dabei nur angrenzende Landschaften oder durch Markierungen verbundene Regionen angreifen. Durch die Verteilung der Armeen auf dem Spielplan bieten sich also jedem Spieler mehr oder weniger viele Angriffsmöglichkeiten. Das Glücken eines Angriffs ist dabei je nach dem Verhältnis der Armeenstärke mehr oder weniger wahrscheinlich.
Das Spielfeld kann, wie in Abbildung 2.7b, auch als Graph dargestellt werden. Hier sind die Regionen die Knoten, die Angriffs- und Bewegungsmöglichkeiten sind die Kanten. Im Sinne von Netzen bilden sich somit Cluster, die von bestimmten Positionen aus angegriffen werden können.
Kontinente sind untereinander nur an speziellen Punkten zugänglich. Alle angrenzenden Regionen oder durch Linien verbundenen Regionen bieten dabei für eine Armee Angriffs- und Zugmöglichkeiten. Diese erreichbaren Regionen sind einfach ersichtlich. Dabei tut sich nur eine Ausnahme auf: Die Meerenge zwischen Asien und Nordamerika. Diese endet am einen Spielfeldrand und fängt am anderen wieder an. Das kann ein unaufmerksamer Spieler übersehen.
Hier zeigt sich der Vorteil von verschiedenen Visualisierungsmethoden. Die geographische Darstellung macht es schwer, die Verbindung zwischen Asien und Nordamerika zu erkennen wie in Abbildung 2.7a. In Abbildung 2.7b ist die Verbindung (am Oberen Rand der Darstellung) und ihre Wichtigkeit einfach ersichtlich. In spielen wie Risiko ist das Spielfeld Vorgegeben anders als bei Trans-America. Hier müssen die Spieler gemeinsam ein Schienennetz durch Nord-Amerika bauen und als Ziel Städte verbinden. Alle Spieler können auch das von Konkurrenten gebaute Schienennetz nutzen und an das bestehende Netz anbauen. Dadurch ergeben sich kooperative Dynamiken im Spielverlauf. Auch das Herrausbilden eines Schienennetzes auf der Grundlage des Spielfelds und der Züge aller Spieler ist sehr dynamisch.[68] Die Darestellung des Spielfelds angelehnt an geographische

Abbildung 2.7: *Verschiedene Ansichten des Risiko Spielfelds*



(a) *Risiko Spielfeld als geografische Karte[191]*



(b) *Risiko Spielfeld als Netzwerk[92]*

bedingungen ist bei Trans-Amerika unproblematischer als bei Risiko, da das Spielfeld an allen enden begrenz ist und nicht, wie bei Risiko, an der anderen Seite wieder anfängt.
Auch wenn beide Spiele als geographische Netze betrachtet werden können sind die Spiel"-mecha"-nismen und die Charakteristika einer netzbetrachtung grundlegend unterschiedlich. Diese Aspekte definieren viele Facetten eine betrachtung eines sachverhalts aus Netzmodell. Während bei Risiko gut verteigbare Positionen strategisch wichtig sind, geht es in Trans-Amerika um kürzeste Pfad die teilkooperativ gebaut werden.

2.2.5 Systeme, Simulation und KI

Bei Netzen handelt es sich auch um Abbildungen von Systemen. Dabei ist gerade die Nicht-Linearität von Interesse.
In Netzen passieren Dinge parallel; dies ist beispielsweise in der Agent-Based Simulation der Fall. Kleine Teilroutinen treten hier als Akteure auf. Diese Akteure stehen in Beziehungen zueinander, agieren und beeinflussen sich so gegenseitig. Es ist wie ein Computerspiel, in dem es nur Computer als Spieler gibt.
Es gibt eine Menge Simulationen, die Verkehr simulieren, indem Akteure durch ein Straßennetz geschickt werden.[210, 175] Ähnlich kann auch die Verbreitung von Krankheiten simuliert werden.[90] Simulationen aus diesem letzteren Bereich können helfen, den Verlauf von Epidemien vorauszusagen. Netze bilden hier sich wahlweise aus den Interaktionen zwischen den Menschen. Alternativ werden sie aus einem Ortsnetz und den sich daraus ergebenden möglichen Krankheitsübertragung gebildet. Mögliche Nutzen ergeben sich aus der Entwicklung von Präventionsstrategien.

Petrinetze

Abgesehen von Agenten basierten Systemen können auch Verarbeitungssysteme wie Schaltungen und biologische Prozesse simuliert werden. Im Weiteren wird ein System aus der Simulation vorgestellt, die Petri-Netze. Petri-Netze sind ein Modellierungsansatz aus der theoretischen Informatik. Es handelt sich um ein parallel ausführbares Programm. Die Struktur des Programms ist der Graph.
Bei Petri-Netzen gibt es Stellen und Transitionen. Die Stellen speichern Tokens. Tokens sind weitergegebene Markierungen, die je nach Zeitpunkt des Systems über die Transitionen weitergeleitet werden.
Transitionen bestimmen, auf welche Weise Tokens durch das Petri-Netz "wandern". Eine Transition enthält dabei Regeln, unter welchen bedingungen die Tokens, wann weitergegeben werden. Dies geschieht durch die Gewichtung der gerichteten Kanten. Die eingehenden Kanten zeigen, wie viele Tokens auf der eingehenden Stelle vorhanden seien müssen, um die Transition anzuregen. Die ausgehenden Kanten zeigen mit ihrem Gewicht an, wie viele Tokens von der Transition an die angeschlossenen Stellen, beim Feuern der Transition, ausgegeben werden.
Die Systeme haben auch Quellen und Senken, aus denen Tokens hervorgehen oder die Tokens aufnehmen.

Ein Petri-Netz kann in verschiedener Komplexität vorliegen. In grundlegenden Petri Netzen kann auf jeder Stelle nur ein einzelnes Token liegen, in anderen, komplexeren Petri-Netzen können mehrere Tokens auf einer Stelle liegen. Eine andere Erweiterung ist das Verwenden von verschiedenen Arten von Tokens. Auf diese Weise werden die Möglichkeiten für Transitionen komplexer, beispielsweise wenn sie nur feuern wenn bestimmte Tokens vorliegen. Petri-Netze lassen sich anhand ihrer Erreichbarkeit analysieren. Einzelne Knoten können somit vom Eingang aus erreicht werden, andere nicht. Dies ist eine Eigenschaft eines gerichteten Graphen. Positionen von Transitionen sind aufgrund ihrer strukturellen Eigenschaften bewertbar. Transitionen können tot sein, wenn sie durch das System nicht aktivierbar sind, also niemals feuern. Des Weiteren werden bestimmte Lebendigkeitsstufen unterschieden, je nachdem wie häufig eine Transition feuert. Für die Lebendigkeit einer Transition spielen Kreise und sich daraus ergebende selbsterhaltende Systeme eine Schlüsselrolle.
Pertri-Netze machen es beispielsweise möglich, biologische Prozesse zu modellieren. Vor allem die Nebenläufigkeit der Prozesse ist in dieser Form der Simulation interessant.[82]

Künstliche Intelligenz

In der künstlichen Intelligenz gibt es verschiedene Ansätze, die auf Netzen beruhen. Diese entstehen durch Entscheidungsprozesse und sind teilweise der Vorstellung geschuldet, dass das Gehirn nicht linear, wie beispielsweise ein klassischer von Neumann-Rechner arbeitet. Ein Gehirn wird dabei als massiv paralleles System gesehen.

Neuronale Netze

In unserem Gehirn gibt es Milliarden von Neuronen. Diese Neuronen bilden eine Art Kommunikationsnetz. Sie verbinden sich und schaffen, auf noch ungeklärte Weise, Intelligenz. Ein neuronales Netz, wie es in der künstlichen Intelligenz benutzt wird, ist dabei ein vereinfachter Nachbau dieses noch nicht wirklich fassbaren Prozesses.[163]
Neuronale Netze basieren auf Impuls- oder Flussnetzwerken. Das Eingangssignal von neuronalen Netzen kommt über die Eingangsknoten herein und das Ergebnis wird durch die Ausgangsknoten ausgegeben. Dazwischen befinden sich versteckte Ebenen, sogenannte hidden Layer, von Knoten und Kanten, die als Verarbeitungsschicht dienen. Am Ende des neuronalen Netzes stehen ein oder mehrere Ausgangsknoten, deren Aktivierung ein Ausgangssignal ausgibt.[126, s. 161 ff]
Dieses Minimalmodell eines Gehirns wird durch positive und negative Rückmeldungen konditioniert. Um das Netz zu trainieren werden Eingangsreize angelegt. Diese sollen auf der Ausgangsschicht ein gewünschtes Muster erzeugen. Immer wenn das Netz richtig reagiert wird ein positiver Impuls an das Netz gesendet. Das verfestigt die Bindung der Neuronen. Auf diese Weise werden die Verbindungen und Signalwege geformt.
Genutzt werden solche Systeme u.a. zur Klassifizierung.[21]

[21] Die Firma Google nutzte dieses Verfahren, um Katzen in Bildern und Videos zu erkennen.[255]

Markovnetze

Diese Art von Netzen basiert auf Wahrscheinlichkeiten von Zustandsübergängen. Die Knoten in einem Markovnetz beschreiben dabei einen Zustand. Eine Kante ist ein Zustandsübergang zwischen zwei Zuständen mit der Wahrscheinlichkeit des Übergangs. So entsteht ein lernendes und voraussagendes Netz.
Jede neue Erfahrung eines Zustandsübergangs kann dabei in die bestehenden "Erfahrungen" integriert werden. Aus diesen Übergangswahrscheinlichkeiten ergibt sich ein gerichtetes und gewichtetes Netz.
Die enstehenden Netze können zur Vorhersage von Zuständen oder Abfolgen genutzt werden. Das geschieht, indem die Markovnetze zufällig durchlaufen werden und die dabei entstehenden Pfade sequenziell abspeichert.
Zur Konstruktion dieser Netze können beispielsweise Texte verwendet werden. Dabei formen Worte die Knoten. Die wahrscheinlichkeit der Abfolge von Worten wird durch die Kanten abgebildet. Die Kanten bilden auch die Übergangswahrscheinlichkeiten ab.
Konstruiert werden solche Netze aus bestehenden Texten. Die aufeinander folgende Worte bilden die Kanten. Kommt eine Wortfolge häufiger vor wird die Kante mit einer höheren Wahrscheinlichkeit gewichtet.
Aus diesen Markovnetzen lassen sich dann Texte generieren indem sie zufällig durchlaufen werden und die Abfolge Knoten entspricht den aneinander gereihten Worten.[123]

2.3 Das Netzwerk als Modell

Im vorherigen Abschnitt wurden verschiedene Anwendungen und Untersuchungen, die auf Netzen beruhen und für die Geisteswissenschaften interessant sind, vorgestellt. Die theoretische Graphenmodellierung bietet Möglichkeiten, ein mehr oder weniger komplexes Modell auf zu stellen. Dieses Modell wird Ergebnisse von Arbeitsprozessen und Datenerhebungen abbilden.
Dabei sind Abbildugskonzepte nur beschränkt ineinander überführbar. Wie die Beispiele zeigen, können nicht alle Fragestellungen können an einem Belibigen Graphen untersucht werden. Für bestimmte Möglichkeiten der Verarbeitung sind spezielle Eigenschaften nöting, die in den Daten enthalten sein müssen. Für die Visualisierung ist dabei ist zu beachten, dass ein komplexes Modell mehr Symbole und Eigenschaften verwenden muss um komplett dargestellt zu werden. Hier werden auswahlkriterien wichtig.
Im Folgenden werden die wichtigsten Eigenschaften zur Graphenmodellierung vorgestellt. Im Rahmen dessen werden Visualisierungs- und Verarbeitungsmöglichkeiten aufgezeigt. Auch werden Möglichkeiten gezeigt, wie Eigenschaften in andere Eigenschaften überführt werden können. Das hilft im Endeffekt, die Daten auf eine Problemstellung hin zu untersuchen.

2.3.1 Richtung

Gerichtete Kanten haben immer einen Ausgangs- und einen Zielknoten.
Ungerichtete Kanten können im Gegensatz zu gerichteten Kanten als ungeordnete Menge von zwei Knoten gesehen werden. Wichtig ist dabei der Kontext, in dem ein Graph verwendet wird.
Hat ein Knoten in einem gerichteten Graphen keine Eingangskanten, dann wird er auch als **Quelle** bezeichnet. Hat er dagegen keine Ausgangskanten wird er als **Senke** bezeichnet. Quellen und Senken stellen häufig Systemgrenzen in Form von Eingängen und Ausgängen dar.
Durch gerichtete Kanten können konkrete Flüsse von Dingen dargestellen, wie Waren- oder Flüssigkeitsströme. Die Richtung von Kanten kann auch als eine Art Diode, Ventil oder Einbahnstraße gesehen werden. Dadurch wird der potentielle austausch und die Richtung beschränkt.
Auch intellektuelle Beeinflussung von Personen untereinander kann als Fluss verstanden werden. Eine Beeinflussung kann von einer Person zur anderen gehen. Eine gerichtete Beziehung kann einen möglichen oder konkreten Einfluss abbilden, wie z.B. die intellektuelle Beeinflussung von Personen untereinander. Diese wäre dann durch die Abfloge der Ideen beschränkt.
Ein anderes Beispiel sind die Abbildungen von Referenzen, wie Links in HTML[22] Dokumenten, dort wird von einer Seite auf eine Andere verwiesen und es geht nicht zwangsweise eine Verbindung zurück.

[22] Siehe Abschnitt 2.2.3.

Gerichtete Kanten bilden bei Personen immer asynchrone Beziehungen ab.
Eine Richtung kann beispielsweise ein Schuldnerverhältnis darstellen. Dabei ist die Quelle der Schuldner und das Ziel der Verleiher, da ein Flusspotenzial an Geld vom Schuldner zum Geldverleiher geht. Die Richtung könnte jedoch auch das Gegenteil bedeuten, der Pfeil geht vom Geldverleiher zum Schuldner, da der Geldverleiher beispielsweise Einfluss auf den Schuldner ausüben kann, da er in der stärkeren Position ist. Die Richtung hängt in diesem Fall von der Fragestellung ab.

Im Allgemeinen können ungerichtete Netzwerke in gerichtete Netzwerke überführt werden, indem eine ungerichtete Kante, durch zwei gerichtete Kanten zwischen den gleichen Knoten ersetzt wird. Beim Umformen von gerichteten zu ungerichteten Kanten stellt sich die Frage, welche allgemeinere symmetrische Beziehung aus einer eigentlich asymmetrischen Beziehung hervorgeht. Als Beispiel kann hier die Untersuchung von Breiger und Pattison dienen[43]. In Padgetts Untersuchungen zu venezianischen Familien[192] wurden Schuldnerverhältnisse nur als ungerichtete Kanten codiert. Diese eigentlich asynchronen Verhältnisse haben Breiger und Pattison dann pragmatischerweise zu Geschäftsverhältnissen uminterpretiert.
Eine Beziehung kann dadurch konkretisiert werden. In einem Verwandtschaftsnetzwerk kann es wichtig sein, in welche Richtungen die Beziehungen laufen, um beispielsweise Vater und Sohn auseinanderzuhalten. Die Darstellung von Vater zu Sohn muss hier mit dem Typen der Kanten zusammenhängen. Diese Art von Beziehung lässt sich auch umgekehrt darstellen, also als Sohn zu Vater. Hier hat sich jetzt jedoch die Bedeutung der Kante verändert. Die Kante bedeutet soviel wie “ist Sohn von”.
Im gleichen Netz kann aber auch eine Bruder-Kante existieren. Diese wäre symmetrisch. Das heißt, hier würde die Richtung keinen Einfluss haben. Wenn jedoch klar ist, wer der ältere Bruder ist und dies in der Abbildung von Belang ist, könnten hier wieder gerichtete typisierte Kanten genutzt werden.
Wenn eine Kante nur ein Verwandtschaftsverhältnis im Generellen zeigt, ist eine Richtung der Kanten nicht nötig.
In der Praxis gibt es oft Varianten von Algorithmen für gerichtete und ungerichtete Verbindungen. Dabei spielt es auch oft eine Rolle, ob ein Algorithmus einen Graphen nur mit der Richtung durchlaufen kann oder entgegen der Richtung. Diese Information ist für Pfadsuchen sehr Wichtig.
Bei der Analyse von verbundenen Teilgraphen stehen bei gerichteten Graphen zwei verschiedene Verbundenheiten zur Diskussion, die schwache und starke Verbundenheit.
Die stark verbundenen Teilgraphen sind unter Berücksichtigung der Kantenrichtung erreichbar.
Schwach verbundene Teilgraphen sind unter Vernachlässigung der Richtung erreichbar.

Visualisierung

In Darstellungen von gerichteten Kanten werden die Enden der Kanten oft mit Pfeilen versehen. Der Pfeil geht von der Quelle zum Ziel. Der Pfeilkopf befindet sich am Ziel-

knoten. Andere Darstellungen verwenden statt eines Pfeilkopfes eine sich zum Zielknoten verjüngende Kantenlinie.

2.3.2 Selbstreferenzen

Selbstreferenzen sind Kanten von Knoten zu sich selber. Start- und Zielpunkt sind dabei nicht zwei verschiedene Knoten, sondern ein und derselbe Knoten.
Ein solcher Selbstverweis kann beispielsweise die interne Vernetztheit eines Systems beschreiben, das in einem Knoten zusammengefasst wird.
Dabei kann eine starke Selbstkante z.B. ausdrücken, dass die Beziehungen innerhalb eines Knotens der eine Organisation oder Gruppe von Personen beschreibt, stark ist.
In einem Markovnetz, in dem eine Kante den Übergang von einem Zustand zu einem anderen darstellt, wird eine solche Kante als Möglichkeit des Übergangs zum gleichen Zustand gesehen.
Im Fall einer Selbstreferenz wäre es das Verharren im gleichen Zustand. Im Ausgangstext wäre eine Dopplung in der mündlichen Sprache eine solche, die eine Bekräftigung darstellt, beispielsweise in der Aussage "Ich war sehr sehr müde". Hier gibt es eine Transition vom Wort "sehr" auf das Wort "sehr".

2.3.3 Gewichte

Eine Gewichtung tritt dann auf, wenn Knoten und Kanten mit einem spezifischen numerischen Wert versehen werden. Dies kann Wichtigkeit oder Stärke einer Kante oder eines Knotens bedeuten.

Knotengewichte

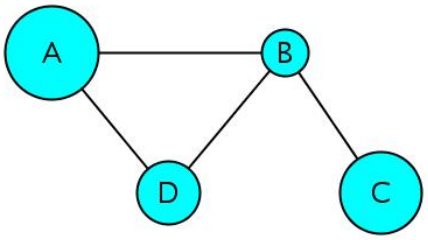
Knotengewichte können sich aus den Eigenschaften des modellierten Sachverhalts ergeben. Diese Eigenschaften lassen sich dem Knoten einfach zuordnen. Hier können Werte von mindestens ordinalem Skalenniveau[23] angenommen werden. Aber auch kontinuierliche Werte können dargestellt werden.
Wenn Knoten beispielsweise Städte und Siedlungen repräsentieren, dann wäre ein Gewicht beispielsweise die Bevölkerungszahl, die Fläche, die Siedlungsdichte, das Vermögen der Einwohner oder das Handelsvolumen.
Im Beispiel von Personen können dies einfache Eigenschaften wie das Einkommen, das Gewicht oder die Bildung sein.

Visualisierung

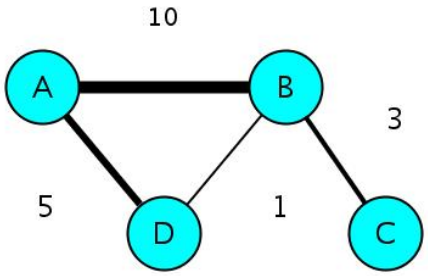
In der Visualisierung können Gewichte durch die Größe eines Knotens dargestellt werden. Größere Knoten haben größere Kreisdurchmesser, Flächeninhalte oder Symbole ein höheres Gewicht. Das ist in Abbildung 2.8a dargestellt. Eine andere Darstellungsweise ist die Abbildung des Gewichts durch Farben oder Farbintensitäten.

[23] Ordinales Skalenniveau im Sinne einer hierarchischen Ordnung von Werten.

Abbildung 2.8: *Gewichtung von Knoten und Kanten*



(a) *Darstellung eines simplen Beispiel-Graphen mit gewichteten Knoten.*



(b) *Beispiel-Graph mit gewichteten Kanten. Die Kantengewichte in Zahlen stehen an den Kanten.*

Kantengewichte

Gewichtete Kanten sind eine mathematisch sehr dankbare Eigenschaft. Viele Modelle der SNA sind darauf ausgelegt, mit Gewichten zu arbeiten. Verfügbare Algorithmen können häufig Gewichte von Kanten einbeziehen.
Gewichte können auch zum Senken der Komplexität dienen. Kanten mit sehr geringen Gewichten können je nach Vorgehen herausfiltert werden. Dieses Verfahren ist bei sehr dichten Netzen und vor allem deren Visualisierungen ein Mittel um Komplexität abzubauen und unwichtige, hier schwache, Relationen heraus zu filtern. Im weiteren kann bei automatischen Extraktionsmethoden das Problem auftreten, dass eine Kante ein Gewicht von null besitzt. Hier ist zu prüfen, ob eine Kante mit dem Gewicht null überhaupt eine Kante ist oder ob sie weggelassen werden sollte.
Ein Kantengewicht kann verschiedene Aussagen haben. Als nackte Zahl ist ein Gewicht nicht sehr Informativ; meist wird angenommen, dass es ein Maß für Nähe und Intensität ist. Dabei ist eine hohe Ausprägung als große Nähe oder Stärke einer Beziehung zu sehen. Wenn jedoch die geografische Entfernung auf einer solchen Kante liegt, dann hat sich der Sinn umgekehrt. Das heißt: Umso höher der Wert auf einer Kante ist, umso aufwendiger ist es, sie zu überwinden.
In der Elektrotechnik gibt es eine ähnliche Unterscheidung, wenn von Widerstand und Leitwert gesprochen wird. Diese beiden Maße lassen sich auch einfach — durch Invertierung — ineinander umrechnen.

Visualisierung

Häufig wird die Dicke von Kanten gewählt, um ein Gewicht abzubilden. Eine dicke Kante zeigt dabei meist einen positiven Zusammenhang an. Die Ausprägung der Kantendicke ist dabei abhängig von der Varianz der Ausprägung des Kantengewichts. Ein Beispiel dafür finden wir in Abbildung 2.8b. Es kann aber auch eine Farbe, ein Farbverlauf oder Transparenz gewählt werden, um das Kantengewicht abzubilden.

Bei der Positionierung der Knoten wird ein hohes Kantengewicht meist als starke Anziehung oder Argument für eine nahe Positionierung der verbundenen Knoten interpretiert.

2.3.4 Typisierte Graphen

Ein typisierter oder auch gefärbter Graph, englisch "labled graph" ist ein Graph, dessen Knoten und/oder Kanten typisiert sind. Diese Typisierungen müssen nicht abhängig voneinander sein können jedoch auch Komplexen regeln folgen wie bei semantisch ausgezeichneten Graphen[24]. Typisierte Kanten beschreiben einen Bedeutungszusammenhang oder eine spezielle Beziehung. Es kann auch von einer Klassifikation gesprochen werden.
Im Gegensatz zur Gewichtung handelt es sich hier um Variablen auf einem nominalen Skalenniveau. Die Ausprägungen der Typen lassen sich nicht in eine objektive Ordnung bringen.
Die Verarbeitung dieser kategorialen Eigenschaften ist in Algorithmen nicht so verbreitet wie das Einbeziehen der Richtung und des Gewichts. Diese Eigenschaften können jedoch zum Filtern verwendet werden. Auch können einzelne Beziehungen in Gewichte verschiedener Stärke überführt werden.

Darstellung

Die Darstellung von typisierten Knoten und Kanten kann bei wenigen Kategorien über die Farbe erfolgen. Eine weitere Möglichkeit besteht darin die typisierten Kante durch Text zu benennen. Dabei werden bei vielen Kanten große Mengen kurzer Textabschnitte in der Darstellung entsteht. Bei größeren Graphen ist das sehr unübersichtlich, wenn sich beispielsweise beschriftete Kanten überschneiden. Alternativ kann auch ein Knoten eingeführt werden. Dann steht der Typ der Beziehung als Text im Knoten. Dadurch wird für jede typisierte Kante eine weitere Kante und ein Knoten hinzugefügt.[25]

2.3.5 Parallele Kanten

Parallele Kanten verbinden die gleichen Knoten miteinander. Das kann durch die Bedingung verschärft werden, dass sie die Knoten auch auf die gleiche Weise verbinden wie bei typisierten Graphen. Bei gerichteten Graphen werden Kanten, die in entgegengesetzte Richtung laufen, nicht als parallel angesehen.
Diese zwei Bedingungen können dabei unterschiedlich betrachtet werden. Hier muss auch klar von redundanter Information unterschieden werden. Dabei können parallele Kanten, je nach Modellierung, auch redundant sein.
Diese Modellvariante ist oft nicht gut in den hier besprochenen Softwaresystemen abgebildet.

---

[24] Zu nennen sind Systeme wie OWL, welche eine hierarchische Klassifikation von Knoten und Kantentypen zulassen. Wobei auch aber keine Ordnung nach einer Wertigkeit vorgenommen wird.

[25] Diese Darstellungsweise wird beispielsweise vom RelFinder verwendet, siehe Abschnitt 5.2.2.

Verarbeitung paralleler Kanten

Parallele Kanten werden in der Praxis oft in eine gewichtete Kante zusammengezogen. Ob und wie das in einem System Sinn macht, ist im Einzelfall zu klären.
Ein Fall, in dem sich parallele Kanten nicht verstärken, kann in der Elektrotechnik beobachtet werden.
Wenn in einer Gleichstromschaltung zwei gleiche Widerstände parallelgeschaltet werden, dann wird der Gesamtwiderstand niedriger als die parallel geschalteten Einzelwiderstände. Diese parallelen Widerstandskanten haben dann zusammen weniger Widerstand als eine einzelne Widerstandskante.
Eine Kombination auf Basis des Gewichts ist immer abhängig von der Bedeutung des Gewichts als Maß von Distanz- oder Näheelement. Komplexer wird es, wenn es sich um kategorisch bewertete Beziehungen, also tyisierte Beziehungen, handelt. Hier können keine klaren mathematischen Komplexitätsreduktionen vorgenommen werden wie das Aufaddieren der Werte.
Wenn beispielsweise ein Verwandtschafts- und ein Arbeitsverhältnis vorliegt. Es besteht ein Verwandtschaftsverhältnis, z.B. zwischen Brüdern. Der eine Bruder ist der Chef des anderen. Dies kann die Interaktion der Beziehung ändern. Aus zwei Rollen wird eine dritte, die anderen Regeln folgt.
In der Soziologie wäre das mit einer Rollenkombination[168] vergleichbar. Es würde also adhoc eine neue Vereinigungskategorie “Chef und Bruder” eingeführt.

2.3.6 Bipartit und Multipartit

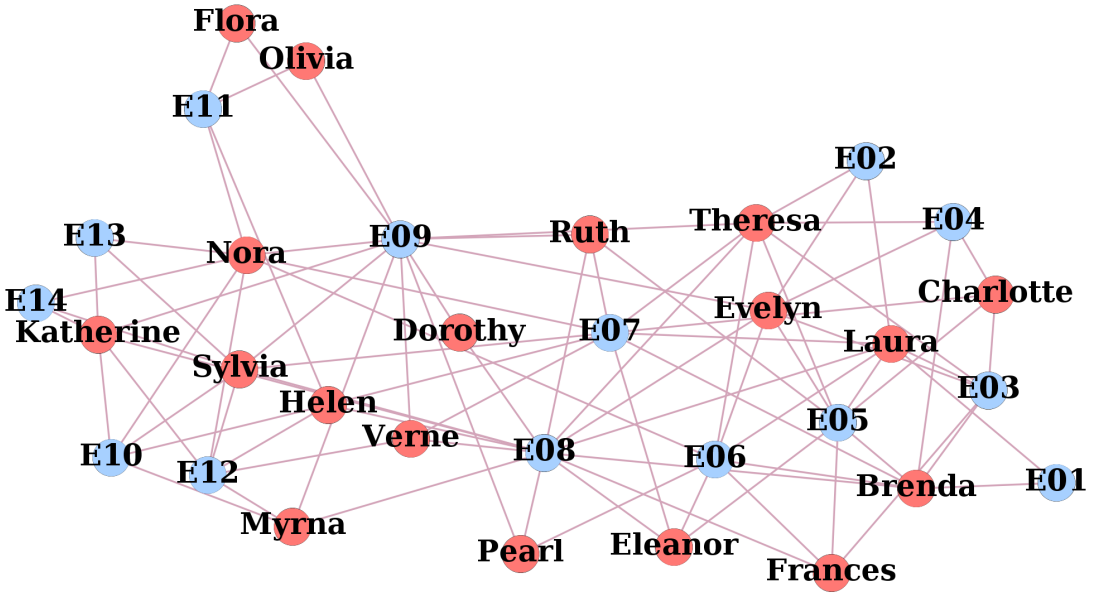
Bipartite Netze, Englisch Two-Mode-Network, zeichnen sich durch zwei Typen von Knoten aus. Die zwei Typen von Knoten überschneiden sich nicht, daher kann ein Knoten nur von einem Typen sein und er kann nicht beide Typen haben. Es sind also zwei disjunkte Mengen von Knoten. Diese Knoten haben keine Verbindungen untereinander.
In einem sexuellen Netzwerk kann es zwei Arten von Geschlechtern geben.[26] Die Relationen bedeuten in diesem Fall sexuellen Kontakt.[88] Im Beispiel kann man aber weitere Fragen an ein bipartites Netz stellen. Die allgemeine Definition sieht bei einem bipartiten Netzwerk nur Kanten zwischen Knoten der beiden Mengen vor. In der realen Welt kann es aber dabei auch vorkommen, dass es Kanten zwischen Knoten in der gleichen Menge gibt. Hier wären es homosexuelle Kontakte, welche die strenge Biparität eines sexuellen Netzwerks durchbrechen würde. Dieser Konflikt kann auf verschiedene Weise gelöst werden. Die homosexuelle Kante muss herausfallen oder es wird auf den Vorteil verzichtet, den ein bipartites Netzwerk für die Analyse bedeutet.
Das klassische Beispiel für ein bipartites Netzwerk ist das Beispiel von Breiger über die Southern Women. Breigers Ziel war es, etwas über die sozialen Kontakte von Frauen untereinander herauszufinden. Direkt war dies nicht zu erfragen. Hier bestehen die Sets der Knoten aus Veranstaltungen und Frauen, die an den Veranstaltungen teilgenommen

[26] Männer und Frauen. Gendertheoretische Überlegungen, dass es mehr als zwei Geschlechter gibt oder diese sich nicht disjunkt ausschließen sind hier der Einfachheit halber außen vor gelassen.

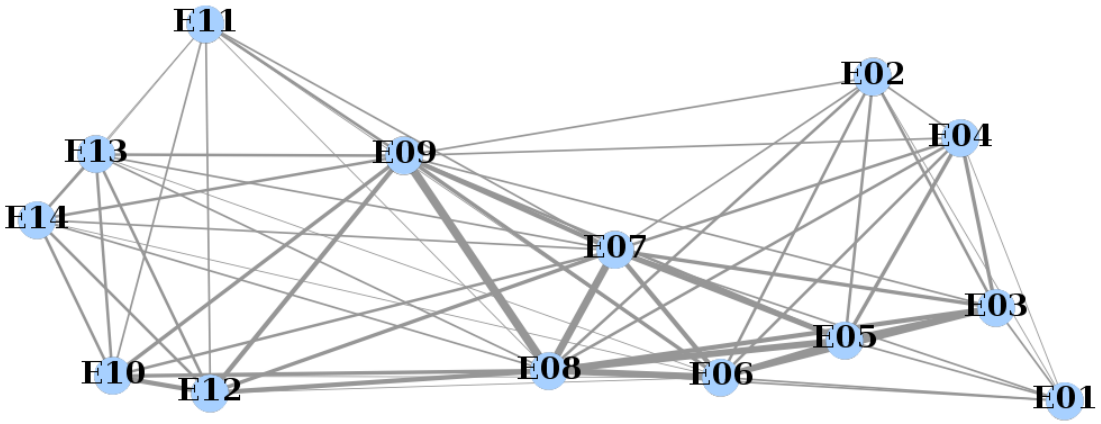
haben. Eine Frau ist mit einer Veranstaltung verbunden, wenn sie diese Veranstaltung besucht hat.[42]
In Darstellung 2.9a [27] sind die Knoten als andersartig gekennzeichnet. Das kann durch Form oder Farbe realisiert werden. Die Andersartigkeit ist meist von kategorialer Art, daher bieten sich Unterscheidungen durch Form und Farbe eher an als durch Größe. Bei der Größe liegt visuell eher eine quantitative Wertung vor als bei einer Farbe oder einer Form.
Bipartite Netze können in einfache gewichtete Netze umgewandelt werden. Dieser Vorgang wird Projizieren genannt. Eine Knotenmenge fällt weg und es ensteht ein unipartites Netz, dessen Kanten auf verschiedene Arten gewichtet werden können. Das Ergebnis einer solchen Projektion wird in Abbilung 2.9b für die Veranstaltungen und in Abbildung 2.9c für die Frauen gezeigt. Das in dieser Arbeit am verwendete listenbasierte Verfahren wird in Abschnitt 3.5.1 beleuchtet. Es gibt aber auch andere Verfahren die auf Matrizen beruhen.
Weitere Beispiele für bipartite Graphen sind, wie in Abschnitt 2.2.2 beschrieben, der Erdős-Graph aus mathematischen Publikationen und Menschen oder der Bacon-Graph aus Filmen und Menschen. Ein multipartiter Graph wäre dabei der Erdős-Bacon-Graph mit Menschen, Filmen und mathematischen Publikationen. Hier ist der Mensch die zentrale Kategorie. Abbilden ließe sich einiges an weiteren Referenzen beispielsweise ein Film, der auf einer mathematischen Publikation beruht oder eine mathematische Publikation, die einen Film zitiert. Die Parität eines Graphen kann auch algorithmisch, ohne das explizite Wissen über die Klassenzugehörigkeit festgestellt werden, jedoch nur bei einem streng bipartiten Graphen ohne Referenzen zwischen den Knoten der gleichen Klasse.
Multipatite oder Multi-Mode-Netze sind Netze, die mehrere disjunkte Knotenklassen aufweisen. Diese Netze sind in ihrer Beschaffenheit komplex. Dies kann beispielsweise für relationale Datenbanken gelten, wie in Kapitel 3 weiter ausgeführt wird.

[27] Datenquelle: `http://vlado.fmf.uni-lj.si/pub/networks/data/GBM/davis.htm`[194] Pajek, Visualisierung Gephi.

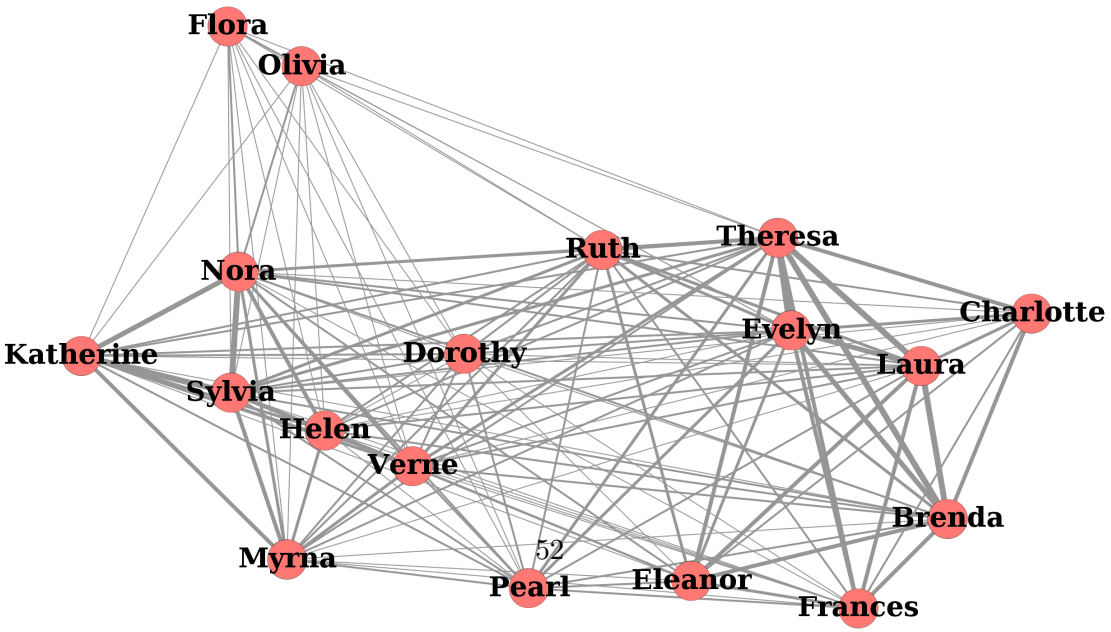
Abbildung 2.9: *Eine Visualisierung des Southern Women Graphen nach den Daten von Homans 1950[194]. Die roten Knoten repräsentieren die Frauen, die blauen Knoten die Veranstaltungen an denen sie teilnahmen.*



(a) Bipartiter Graph aus Veranstaltungen und Frauen.



(b) Projektion der Veranstaltungen über die Frauen



2.3.7 Zeitgraphen

Netzwerke sind meist nicht statisch und ändern ihre Struktur mit der Zeit. Graphen die eine Zeitebene berücksichtigen werden Zeitgraphen oder englisch time-graphs genannt.
Kanten und Knoten können zu bestimmten Zeitpunkten entstehen, verschwinden oder sich und ihre Verbindungen verändern. Das Modell wird hier um die Lebensdauer, das erste Auftreten, das Ende oder die Veränderung einer Kante erweitert. Somit erhält das Modell des Graphen eine Zustandsbeschreibung auf der linearen Zeitachse.
Die Abbildung eines solchen Vorgehens im Computer kann durch die zeitliche Speicherung der Lebensdauer einer Kante bestehen. Auf diese Weise können die Handelsbeziehung zwischen Ländern zu einem bestimmten Zeitpunkt bestehen und dann in einer späteren Zeitspanne verschwinden.
Ein weiteres Beispiel findet sich in der Entwicklung der Beziehungen vor dem Ersten Weltkrieg. Die Bündnisse zwischen den Staaten entwickelten sich dabei über die Zeit hinweg[28].
Die Bündnisse können zu verschiedenen Zeitpunkten abgebildet werden. Daraus ergibt sich ein Zeitgraphen, in dem die Knoten, also die Staaten, konstant bleiben, während die Bündnisse, hier die Kanten, eine Lebensdauer haben.
Ein weiteres Beispiel kommt aus dem Feld der sexuellen Netzwerkforschung. Um die Ausbreitung von Geschlechtskrankheiten zu untersuchen, ist es wichtig den Zeitpunkt einer Interaktion zu kennen.[157]

Visualisierung

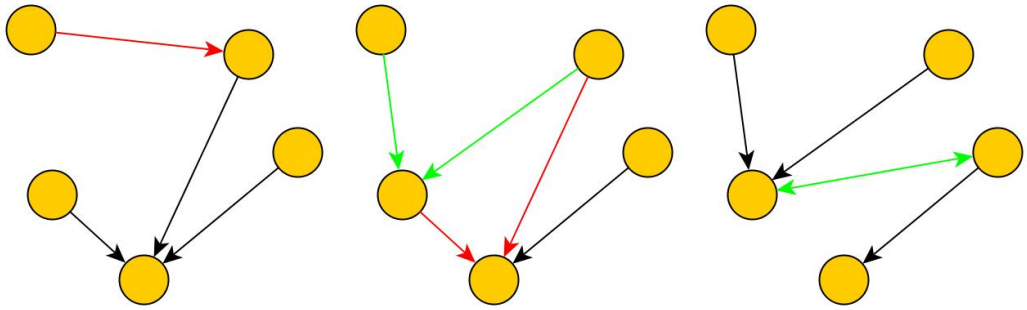
Die Darstellung solcher Vorgänge kann dabei meist nicht aus einem Bild bestehen. Oft wird auch mit Videos gearbeitet, wie bei Gource[57, 52] [29] Es sind aber auch Betrachtungen zu verschiedenen Zeitpunkten möglich, wie beispielhaft in Abbildung 2.10.
Ein weiteres Beispiel für die Abbildung eines Time-Graphs ist die Untersuchung von Schich et. al.[220] zu den Geburts- und Sterbeorten berühmter Menschen. Mithilfe dieser Daten wird ein geografischer Zeitgraphen aufgespannt.
Die Frage, ob ein Zeitgraph gut in einem Video visualisiert werden kann, hängt auch davon ab, wie die Zeit eingeteilt ist. Die Zeit dient dabei als objektive Einteilung des Ablaufs. Sie kann im Vielfachen der realen Zeit ablaufen. Dies kann manchmal nicht sinnvoll sein, da sich in manchen Zeiten mehr verändert als in anderen. In Videoplayern gehört die Steuerung des Ablaufs durch langsames oder schnelleres Abspielen zu den Grundfunktionen, was diesen Punkt etwas egalisiert.

---

[28] Siehe Abschnitt 2.2.1 .

[29] Grouce ist eine Software, die den Ablauf eines Programmiervorhabens in einer Versionsverwaltung darstellt. Die zeitlichen Abläufe werden dabei anhand der Informationen aus der Versionsverwaltung dynamisch visualisiert.[52] Auch die Visualisierungssoftware Gephi, bietet die Möglichkeit Zeitabläufe in Graphen darzustellen.

Abbildung 2.10: *Vorschlag für die Visualisierung eines Zeitgraphen als Serie von Bildern. Dabei sind die neu erscheinenden Kanten grün und die im nächsten Schritt wegfallenden Kanten rot markiert.*



(a) *Zeitpunkt 1* ***(b)*** *Zeitpunkt 2* ***(c)*** *Zeitpunkt 3*

2.3.8 Hypergraphs

Hypergraphs beschreiben Graphen, bei denen die Kanten nicht nur zwei Knoten verbinden können. Die Kanten bilden hier eine Menge, die nicht nur aus maximal zwei Knoten bestehen kann, sondern aus vielen. Es bilden sich Mengen von Knoten, die als Hyperkanten engl. "hyper edges" bezeichnet werden.
Hypergraphen sind meist nicht einfach zu berechnen und es gibt weniger Algorithmen, um diese zu verarbeiten. Daher werden Hypergraphen gerne in einfachere Formen von Graphen überführt, um diese bearbeiten zu können.

2.3.9 Weitere Werte

Prinzipiell können jegliche Daten mit Knoten und Kanten assoziiert werden. Oft ergibt sich eine mögliche Bewertung oder Erkenntnis erst aus diesen Eigenschaften. Es ist nicht direkt Teil eines theoretischen Graphenmodells.
Diese Werte können je nach Modell anders zugeordnet werden. Ein Knoten, der einen Menschen repräsentiert, kann auch dessen Eigenschaften repräsentieren wie Geschlecht, Alter, Status, Lieblingsfarbe, bevorzugtes Reimschema etc.. In einem Straßennetz von Städten fänden hier die Einwohnerzahl, der Name, der Herrschaftsbereich etc. Platz. Bei Büchern wären das z.B. Publikationsdaten, Verkaufszahlen, Thema, Stichwörter.
In der informatischen Praxis besteht eine starke Verbindung zum objektorientierten Programmierung. Knoten und Kanten durch Programmobjekte beliebiger Komplexität repräsentiert werden. Hier gibt es einen starken Zusammenhang zu Gewichten und Farben in Graphen. Je nach Auswertung können diese Daten als Gewichte oder Farben für Verfahren auf Graphen verwendet werden, wie beispielsweise die Errechnung des kürzesten Pfades.

2.3.10 Pfad

Ein Pfad durch einen Graphen ist eine Abfolge von Knoten und Kanten. Dabei kommt in vielen Anwendungen die Einschränkung hinzu, dass alle Knoten und Kanten nur einmal

vorkommen dürfen. Die Beschreibung kann dabei in den meisten Graph-Modellen durch Knoten oder Kanten erfolgen, ausgenommen sind beispielsweise Graphen mit parallelen Kanten. Ein Pfad kann also als eine geordnete Menge von Knoten oder Kanten dargestellt werden.
Das einmalige Vorkommen einer Kante auf einem Pfad ist beispielsweise im Sieben-Brücken-von-Königsberg Problem von Leonard Euler abgebildet. Das Problem besteht darin, einen Pfad durch Königsberg zu finden, der nur einmal über alle Brücken führt. Start und Endpunkt müssen übereinstimmen. Euler fand heraus, dass ein solcher Spaziergang nicht möglich ist.[184, s.19 ff]
Ein daraus hervorgehendes Problem, das sich mit Pfaden beschäftigt und sich vom Sieben-Brücken Problem ableitet, ist das Problem des Handlungsreisenden[30]. Die Aufgabe ist: Ein Handelsreisender soll auf seiner Tour alle Städte besuchen und an seinem Ausgangspunkt auch wieder ankommen. Dabei solle er die kürzeste Strecke wählen. Dieses Problem kann heute bei großen Netzen mit Computern nicht zeitkritisch gelöst werden.[184, s.52 ff] In der Praxis versucht man, durch Annäherung und Heuristiken dem Problem Herr zu werden.
Bei Pfaden ist jeweils auch die Richtung der Kanten zu beachten. Eine Kante mit einer Richtungsanzeige kann dabei in der Regel nur in die eine Richtung durchlaufen werden.

Kürzester Pfad

Eine wichtige Spezifizierung eines Pfades ist der kürzeste Pfad. Als kürzesten Pfad bezeichnet man einen Pfad, der zwei Knoten in einem Netz auf dem kürzesten Weg verbindet. Dabei sind meist die Gewichte oder die Qualitäten der Kanten entscheidend. Das kann am Orbis Modell[31] praktisch betrachtet werden. Eine Kante und ihr Gewicht hängen dabei von der Jahreszeit und anderen Bewertungskriterien ab. Dabei stellt sich die Frage, ob es ein Weg minimaler Länge, der kostengünstigste Weg oder der schnellste Weg sein soll.

Visualisierung

Pfade können auf der Darstellung eines Graphen farblich hervorgehoben werden. Pfade werden auch mit parallel laufenden Bögen[32] hervorgehoben. Diese Abbildungsvariante wurde in Abbildung 2.2 verwendet.

2.3.11 Strukturelle Einschränkungen

Strukturelle Einschränkungen im Graphenmodell können einen Graphen in eine simplere Datenstruktur überführen. In diesen einfacheren Datenstrukturen können manche Informationen nicht abgebildet werden, dafür sind sie einfacher abbildbar und effizienter verarbeitbar.

[30] Engl.: traveling salesman problem

[31] Siehe Abschnitt 2.2.4.

[32] Engl.: arches

Planar

Als planar werden Graphen bezeichnet, die ohne Kantenüberschneidungen abgebildet werden können, das ist bei geografischen Netzen oft der Fall.

Topologie

Eine weitere Einschränkung ist eine Topologie oder DAG[33]. Topologien besitzen keine richtungsgebundenen Zirkel. Das heißt: wird die Topologie in Flussrichtung abgelaufen, dann wird kein Knoten zweimal passiert und jeder Pfad wird in einer endlichen Anzahl von Schritten beendet. Eine Topologie kann aber im Gegensatz zu einem Baum mehrere Knoten ohne Eingangskanten haben.

Bäume

Ein Baum ist eine Einschränkung der Struktur eines gerichteten Graphen. Ein Baum liegt vor, wenn ein Graph zirkelfrei ist und eine einzige Wurzel besitzt. Dabei kann jeder Knoten höchstens eine eingehende Kante besitzen. Besitzt ein Knoten keine eingehende Kante, dann ist das die Wurzel des Baums.
Bäume sind in der Informatik ein gut untersuchtes Feld und werden daher gerne verwendet.
Die Wurzel eines Baumes nimmt dabei die zentrale Position ein. Mehrere Bäume können dabei auch als Wald bezeichnet werden. Dabei kann dieser Wald auch zu einem Baum zusammengefasst werden. Dazu muss eine gemeinsame Wurzel definiert werden.
Als Blätter werden die Knoten bezeichnet, die auf keinen weiteren Knoten verweisen. Beim Durchlaufen eines Baumes endet ein Weg immer an einem Blatt. In Entscheidungsbäumen bilden die Blätter eine Entscheidung ab, während alle anderen Knoten in einem Entscheidungsbaum Fragen abbilden, die zu einem Blatt, d.h. einer Entscheidung, führen.[17, s. 8 ff]
Bäume können auch als Subgraphen extrahiert werden. Ein Beispiel hierfür ist der minimal spanning tree. Er verbindet alle Knoten eines vollen Graphen mit einer minimalen Anzahl von Kanten. Dabei werden alle Zirkel entfernt.[50, s. 101]Im Rahmen einer Modellbildung können Bäume einfache oder idealisierte Strukturen abbilden, wie beispielhaft in Abbildung 2.2.1.

2.3.12 Analytische Strukturen

Es gibt kleinste Graphen, die aus wenigen Knoten und Kanten bestehen. Teilweise sind diese Strukturen wichtig für die Beschreibung von Analysemöglichkeiten und der Definition von Kennzahlen. Kleinstgraphen lassen sich dabei einfach auflisten, da es nicht viele Zustände gibt die sie einnehmen können, beispielsweise alle Kombinationen von Graphen mit fünf Knoten.[96]

[33] Ein gerichteter zirkelfreier Graph. Engl.: directed acylic graph.

Vollständige Cluster

Ein vollständiges Cluster ist eine Menge von Knoten, die mit allen anderen Knoten verbunden sind.
Diese Strukturen können in sehr dichten Graphen auftreten. Vollständige Cluster treten häufig in Projektionen von bipartiten Graphen auf. Diese Teilnetze sind meistens nicht erwünscht und geben hinweise auf Fehler in den Daten.

Zirkel und Dreiecke

Zirkel oder Kreise in Netzen treten genau dann auf, wenn ein Pfad durch einen Graphen seinen Startknoten auch als Endknoten aufweisen kann.
Der einfachste Kreis ist ein Dreieck von ungerichteten Kanten.
Das Fehlen von Zirkeln kann die Komplexität von Berechnungen auf Graphen einschränken. Dies ist beispielsweise bei Bäumen, Wäldern und Topologien der Fall.
Des Weiteren können Zirkel schwach und stark sein, diese Eigenschaft bezieht sich darauf, ob die Richtung der Kanten miteinbezogen wird oder nicht.
Dreiecke sind aber nicht nur die kleinsten Kreise, die in einem Netz Auftreten können, sondern auch die kleinsten Cluster. Die Menge der Dreiecke findet sich im Clusteringkoeffizient[34] wieder. Es gibt auch Theorien über Dreiecksbeziehungen verschiedener Färbung, wie die Balance Theorie[35].
Dreiecke sind recht kostengünstig zu berechnen. Daher bieten sich dreiecksbasierte Analysen auf großen Datenmengen an.
In bipartiten Netzen gibt es keine Dreiecke. Der kleinste Kreis, in einem bipartiten Netz ist ein Viereck. Dieses Viereck besteht dabei aus jeweils zwei Knoten beider Knotentypen.

Stern

Eine Stern-Topologie liegt vor, wenn eine Menge von Knoten mit einem einzigen zentralen Knoten verbunden sind und nicht untereinander. Dies ist ein Muster, das beispielsweise bei bipartiten Graphen vorliegen kann. Der Graph besteht aus einem einzigen Knoten des ersten Typs und mehreren Knoten des zweiten Typs. Da die Knoten des zweiten Typs nicht untereinander verbunden sein können, ergibt sich ein Stern. Dieses Muster bildet sich beispielsweise bei nicht überlappenden Klassifizierungen, die als bipartiter Graph abgebildet werden. Das Zentrum ist die Klasse, um die sich die klassifizierten zugeordneten Entitäten anordnen. Wird ein solches Stern-Muster projiziert, ergibt sich ein vollständig verbundenes Cluster.

[34] Siehe Abschnitt 2.5.5.

[35] Siehe Abschnitt 2.2.1 und 4.3.2.

2.3.13 Von den Daten zu Knoten und Kanten

Ein wichtiger Schritt in jeder Datenmodellierung ist die Identifikation von Entitäten in den Daten. Dabei gibt es einige Fallstricke, die im Fall von Netzen besondere Auswirkungen haben.

Granularität

Granularität ist die Betrachtungsgenauigkeit von Daten. Diese ist für das Vergleichen von Daten wichtig.
Granularität kann an der Analyse von Websites erklärt werden. Die Betrachtung kann auf Ebene der einzelnen HTML-Dokumente und ihrer Links untereinander angesetzt werden. Alternativ kann sie an höherer Stelle ansetzen. Hier wären komplette Websites oder sogar ganze Toplevel Domains zu nennen. Im ersten Fall wird das HTML-Dokument als Knoten abgebildet. Im zweiten Fall wird die Seite bzw. die Domäne als ein Knoten verstanden.
Probleme mit der Granularität können bei schwer fassbaren Mengen wie inoffiziellen Gruppen und Mehrfachzuorndung von Entitäten, zu Problemen führen. Es geht dabei nicht um das direkte Vergleichen von Äpfeln und Birnen, sondern eher um den Vergleich vom Apfel mit dem Apfelbaum oder dem Ast eines Apfelbaums.
Die Änderung der Granularität ändert die Aussage eines Graphen. Die Betrachtungsgenauigkeit und Betrachtungsstufe ist entscheidend für die Stufe der abgeleiteten Aussagen. So ist ein Zusammenfassen einzelner Knoten zu einem weniger granularen Modell ohne Probleme möglich.
In der Soziologie ist das Ganze als Mikro- und Makrophänomen zu betrachten. Die Einflüsse und Handlungen liegen bei den einzelnen Akteuren. Die Konsequenzen dieser parallel laufenden, gleichförmigen Handlungen sind nicht nur auf der Ebene der Personen zu erkennen, sondern lassen sich auch auf der Ebene größerer Zusammenfassungen von Akteuren beobachten. Dieser Effekt wird in Colemans "Badewanne"[36] beschrieben.[56]

Knoten und Identität

Ein weiteres Problem, das sich beim Modellieren eines graphenbasierten Ansatzes ergibt, ist die Identifikation der zu modellierenden Elemente. Dabei ist die Identität von Dingen oft nicht einfach zu bestimmen. Als Beispiel können Namen und Personen dienen. Eine Person kann verschiedene Namen tragen. Das heißt jedoch nicht, dass alle Namen richtig und eindeutig verwendet werden, wenn die Person beispielsweise in einem Text beschrieben wird. Bei ungenügender Datenreinheit und fehlenden Informationen kann es in diesen Fällen zu Verwechselungen kommen, wie bei Personen mit gleichem Namen oder ähnlichen Namen. Umso mehr Kriterien zur verfügung stehen, um Personen auseinanderzuhalten, umso eindeutiger wird die Identität.
Das gilt jedoch nur, solange keine Fehler in den Daten auftreten. Fehler können ansonsten passende Zuordnungen verwerfen. Das Problem tritt in Verbindung mit Daten aus ver-

[36] Engl.: Colemans boat.[56]

schiedenen Quellen auf. Die Quellen haben eigene Bezeichner und einen eigenen Umfang an Daten, die erhoben werden. Beispielsweise wird in einer Quelle der Name und das Geburtsjahr gespeichert, in einer anderen das Geschlecht oder der Geburtsort.
Dabei kann es generell zu zwei Arten von Problemen kommen:
Der Nichtidentifikation von Gleichen:
Es werden Dinge untersucht, die eigentlich das Gleiche sind, jedoch werden sie nicht als das Gleiche identifiziert. Im Modell bilden sich mehrere Knoten, wo eigentlich einer, hätte sein sollen. In der Fachsprache wird auch von falsch negativer Identifikation[37] gesprochen.
Die falsche Identifikation:
Es werden Dinge fälschlicherweise für das Gleiche gehalten, daher werden Informationen zusammengefasst, die nicht zusammengehören, also eine falsch positive Identifikation[38].
Es werden also zwei Personen für die gleiche Person gehalten.
In diesem Fall ist die Nichtidentifikation wahrscheinlich das kleinere Übel, da sie im Nachhinein geändert werden kann. Ein Auseinanderdividieren eines zusammengefassten Knotens ist nach der Kumulierung von Daten nicht immer möglich.
Das Beispiel einer Person ist einleuchtend und einfach, da die normalen Kategorien zur Disambiguation von Personen jedem geläufig sind. Bei anderen Sachverhalten kann das schwerer sein, da die Kriterien der Beschreibung andere sind. Bei Objekten, Kunstwerken, Software, Texten, Ideen, Massenprodukten ist dies nicht mehr ganz so eindeutig. Falsch negativ und Falsch positiv stehen im zusammenhang mit den Begriffen von **Precision** und **Recall**.
Das Precision-Maß gibt das Verhältnis von richtigen Treffern zu falschen Treffern an. Das Maß zeigt an, wie "dreckig" ein Ergebnis ist.
Das Recall-Maß stellt die falsch negativen und wahr positiven Ergebnisse in Verhältnis zueinander. Es zeigt an, wie Vollständig ein Ergebnis ist. Die beiden Maße sind nicht unabhängig voneinander. Wenn es für die Zuordnung von Datenfragmenten zu einer Knoten-Identität mehrere Datenfragmente als Kandidaten gibt und einfach alle dem Knoten zugeordnet werden, auch wenn falsche Zuordnungen darunter sind, dann ist der Recall mit einem Wert von 1 Maximal. Dafür ist das Ergebnis wahrscheinlich auch maximal Dreckig.[201]

2.4 Visualisierung

Visualisierungen von Graphen können hilfreich sein, um Daten zu verstehen. Graphen können viele Informationen auf einmal enthalten. Diese Informationen sind in der linearen Ausgangsform der meisten Daten nicht ersichtlich, die eigentliche Information ist nicht gut linear erfassbar.
Rankingverfahren versuchen eine lineare Ordnung herzustellen. Suchergebnisse werden dabei in einer Ordnung nach ihrer Wichtigkeit in einer Liste dargestellt.
Diese Daten können in textueller und linearer Form viel Platz einnehmen. So kann kein

[37] Engl.: false negative match

[38] Engl.: false positive match

Überblick auf einer Erfassungseinheit also beispielsweise auf einem Bildschirm oder einer Seite geboten werden. Auch kann ein Mensch aufgrund der verhältnismäßig geringen Aufnahmefähigkeit von Einzelinformationen einen Überblick nur durch einen hohen Aufwand erreichen. Hier sollte eine nicht lineare Ordnung gefunden werden, die diesem Fakt Rechnung trägt.
Alternativ kann auch über das Verwenden von Maßzahlen nachgeracht werden. Die Verwendung von Maßzahlen bringt jedoch immer eine starke Komplexitätsreduktion mit sich. Diese Abstraktion ist meist nur für einen Experten, der den dahinter liegenden Mechanismus kennt, klar interpretierbar. Anhand von Anscombes Quartett kann auch gezeigt werden, dass völlig verschiedenen Verteilungen eine ganze Reihe von fast identischen, statistischen Maßzahlen erzeugen.[8] Dem Menschen steht gerade durch seinen Sehsinn ein gutes Werkzeug für die Informationsaufnahme zur Verfügung. Das menschliche Auge kann Informationen schnell gruppieren und in Bruchteilen von Sekunden Strukturen erkennen, die mit dem Computer alleine nur aufwendig zu berechnen sind.
Die Erfassung und Verarbeitung von Daten kann fehlerbehaftet sein. Fehler in Datenmengen lassen sich für Menschen visuell einfach finden, während ein Computer die Überprüfung von Fakten meist nur aufwendig und nur nach bekannten Mustern bewerkstelligen kann. Hier ist Visualisierung auch ein Werkzeug, das dem Menschen hilft, diese Muster zu formulieren.
Im Folgenden werden die gängigsten Formen der Darstellung beschrieben.

2.4.1 Die Matrize

Die Darstellung als Matrize ist ein einfach zu verwendendes Instrument, das auch händisch ohne Computer gut verwendet werden kann. Die Größe der Darstellung hängt von der Menge der Knoten ab. Die Knoten werden auf Zeilen und Spalten abgetragen. Eine Zelle wird mit einem Wert versehen oder einfach gefüllt, wenn die Knoten in Zeile und Spalte miteinander verbunden sind. In dieser Form lassen sich beliebig viele Knoten darstellen. Dabei wächst die Menge der Felder in der Matrize quadratisch an, was die Darstellbarkeit praktisch schon recht früh einschränkt. Die Einschränkungen hängen dabei von der Größe der Darstellungsfläche oder der Matrizenfelder ab. Bei dünn besetzten Graphen ist diese Darstellung zudem problematisch, da ein Großteil der Visualisierungsfläche für nicht existierende Kanten verwendet wird.
Die Matrizendarstellung eines ungerichteten Graphen kann dabei einen ästhetischen Gesichtspunkt haben. Die Matrix lässt sich an der Diagonalen spiegeln. Das vermittelt einen Eindruck von Symmetrie und von Sortiertheit.
Die Matrizendarstellung hat ihren Ursprung in der Mathematik, wo Operation auf Graphen oft mithilfe von Matrizenoperationen beschrieben werden können.
Eine Methode, die aus der Matrizendarstellung hervorgeht, ist das Blockmodelling. Hier werden die Spalten und Reihen einer Matrizendarstel-lung so sortiert, dass sich möglichst viele Blöcke bilden. Diese Blöcke sollen als Nebenbedingung möglichst lückenfrei sein, dh. sie sollen keine freien Positionen im Block haben.
Händisch ist auch diese Aufgabe aufwendig. Der Blockmodellierende muss die Sortierung

finden, die ein möglichst gutes Ergebnis mit möglichst kompletten Blöcken findet. Dabei gibt es viele Permutationen der Anordnung der Knoten auf einer Matrize, die verglichen werden müssen. Diese möglichen Sortierungen lassen sich mit Hilfe von Algorithmen mathematisch einfach bewerten. Es wird geometrisch überprüft, welche Anordnung die Anforderung Blöcke zu bilden am besten erfüllt. Die Vorstellung einer guten visuellen Erfahrung des Menschen, die Darstellung von zusammenhängenden Strukturen als Blöcke lässt sich einfach in die Matrizen-Abbildung überführen. [40, Kapitel 10]

2.4.2 Netzwerk Diagramme

Zweidimensionale Symboldarstellungen bestehen meistens aus einer Menge von Symbolen, oft Kreisen aber auch anderen Symbolen, welche die Knoten repräsentieren. Diese Knoten werden dabei mit den Kanten verbunden.
Graphen werden meistens in zwei Dimensionen visualisiert. Dies ist sicherlich einer gewissen Tradition geschuldet. Karten werden traditionell auf Zwei Dimensionen dargestellt, auch wenn es Dinge wie Höhenkarten oder Globen gibt. Diese nicht zweidimensionalen Visualisierungen sind in der Praxis nur schlecht zu handhaben, verbrauchen viel Platz und sind teuer und aufwendig in der Herstellung. Die Darstellung auf begrenzten zwei Dimensionen ist auch oft einer Tradition der Publikation geschuldet.
In einer Publikation in Buchform ist es nicht möglich, großflächige Darstellungen kostengünstig zu hinterlegen. Auch eine dreidimensionale Darstellung ist in dieser traditionellen Form schwer möglich.
In Abschnitt 2.3 wurden schon einige Darstellungsweisen für einzelne Symbolvarianten aufgeführt. Es gibt für die Darstellung der Symbole an sich einige Probleme wie die Größe oder Codierung der darauf abgebildeten Werte.
Ein anderes Problem ist die Anordnung der Symbole für Knoten und Kanten. Diese Anordnung muss algorithmisch geschehen. Dabei stellen sich die Fragen, wie die Symbole und Verbindungen gut angeordnet werden können und was eine gute Anordnung ausmacht.

Position und Wert

Klassisch werden zweidimensionale Darstellungen oft als die Abbildung von zwei Skalen verstanden.
Dies sind die X- und Y-Achse. Auf diese Achsen werden die Eigenschaften einer Untersuchungsklasse abgebildet. Dadurch werden zwei Eigenschaften ins Verhältnis zueinander gesetzt.
Diese Darstellungsweise benötigt im besten Fall kontinuierliche Werte auf den Achsen. Sie ist damit nicht auf alle Informationen anwendbar und hat die Einschränkung, dass nur zwei Werte zueinander ins Verhältnis gesetzt werden können. Dabei drückt die Position im Raum eine Ähnlichkeit zwischen den auf den Achsen abgetragenen Werten aus.
Diese Abbildung setzt die Dimensionen klar fest. Bei der Abbildung einer Menge von mehr als zwei oder höchstens drei Werten ist diese einfache "Kartierung“ des Problemraums nicht möglich. Natürlich können zusätzlich Symbole, Farben und Größen verwendet

werden. Kontinuierliche Werte haben den Vorteil, dass sie eine lineare Komponente haben, sie steigen von null oder dem Minimum zum Maximum an.
Diese Form der Darstellung setzt dabei die Position als klare Beschreibung fest. Jede Position hat eine eindeutige Bedeutung, solange der Betrachter die Achsenskalierung im Auge hat.
Das kann jedoch eine Menge von Überdeckungen verursachen. Eine Möglichkeit, dem entgegenzuwirken liegt darin, eine Skala nicht linear auf einer Achse abzubilden, sondern logarithmisch. Auch kann die Achse erst ab einem Schwellenwert genutzt werden. Diese Tricks haben den Vorteil, dass der Platz besser genutzt wird, da einzelne Knoten oder Punkte mehr Platz bekommen. Diese Arten der Darstellung beinhalten dabei eine Verzerrung, die der Betrachter sich bewusst sein muss.[133, Kapitel 5]
In den hier beschriebenen Plots ist der Raum gut und objektiv beschreibbar. Ein Verhältnis wird durch die Objektive einteilung global hergestellt.
Die Abbildung der scheint recht eindeutig, doch gibt es hier bei der Übertragung von Werten in den abstrakten Zahlenraum potenziell Probleme.
Menschliche Sinneseindrücke verlaufen nach dem Weber-Fechner-Gesetz im logarithmischen Verhältnis zum physikalischen Reiz. Das Weber-Fechner-Gesetz beschreibt aber auch, welche Unterschiede kognitiv wahrnehmbar sind. Diese Abstände ändern sich auch mit der Intensität des Impulses.[228] Das entspricht der logarithmischen Darstellungsweise in Plots. Als Beispiel kann man hier die Lautstärke dienen. Um die wahrgenommene Lautstärke zu verdoppeln, muss der Schalldruck vervierfacht werden.
andererseits können Übertragungen von Größen auf Symbole irreführend sein. Beispielsweise soll ein Symbol das doppelte eines anderen Symbols darstellen. Dafür wird die Kantenlänge verdoppelt. Dadurch entsteht ein größerer Eindruck als “das doppelte”, denn durch die Verdoppelung der Kantenlänge wird die Fläche vervierfacht.[133, Kapitel 6]
Das Abtragen von objektiven, eindeutigen Werten ist oft nicht so unproblematisch, wie es scheint. Das Problem verschärft sich für die oft nicht gut quantifizier und objektivierbaren Daten aus dem weiteren Umfeld der Geisteswissenschaft.
In assoziativen Abbildungen, in denen kein genereller, sondern ein für jeden Knoten expliziter Zusammenhang herrscht, ist es schwieriger die dargestellten Sachverhalte zu ordnen. Hier kann kein objektives Kriterium herangezogen werden um Sachverhalte in Position zueinander zu setzen. Damit Relationen gut sichtbar sind und das menschliche Auge Strukturen erkennen kann, müssen daher andere Ordnungskriterien genutzt werden.

Diagramme und freie Abbildung

Die freie Abbildung von Knoten und Kanten kann in leicht veränderter Form in Mindmaps, Diagrammen etc. gefunden werden. Sie werden von Menschen erstellt und funktionieren auch bis zu einer gewissen Größe und der Fähigkeit des Menschen.
Das Vorgehen kann dabei von Computern unterstützt werden. Die Ausrichtung von Symbolden kann an einem Gitternetz vorgenommen werden. Das hilft beim Erstellen eines symmetrischen Bildes, da Abstände im Verhältnissen zueinander stehen.
Mindmaps und Diagramme müssen keinen strengen Regeln für die Visualisierung unterlie-

***Abbildung 2.11:** Ausschnitt aus der Tabula Peutingeriana[262] vom "Stiefel" Italiens und Siziliens*

gen. Der Mensch ist für die Sortierung und Anordnung zuständig. Es gibt keine formalen Regeln, nach denen ein Mensch vorgehen muss. Auch sehen nicht alle Realisierungen einer Mindmap durch verschiedene Menschen, gleich aus. Sie stellen daher kein standardisierbares Mittel zur Visualisierung im Sinne der Informatik dar, jedoch sind diese Darstellungen in manchen Fällen als Vorbild zu sehen.

Netzwerkkarten werden bei Interviews eingesetzt und helfen bei der Ordnung des Umfeldes eines Befragten.[242]

Geografische Distanzen

Wenn die Knoten eines Netzes geografische Einheiten repräsentieren, dann liegt es nahe, die Knoten nach ihrer geografischen Position anzuordnen.

Damit eine planare zweidimensionale Fläche als Abbildungsgrundlage verwendet werden kann, muss die runde Erdkugel auf eine Fläche projiziert werden. Dies ist ein gut erforschtes Gebiet und stellt auch algorithmisch keinen problematischen Rechenaufwand dar. Es bleibt jedoch die Herausforderung, eine Darstellung zu finden, die am wenigsten verzerrt. Dabei ist es auch wichtig zu betrachten, ob es sich um die Darstellung von lokal begrenzten Daten oder globalen Daten handelt. Karten, die kleine Ausschnitte der Welt zeigen,

können problemlos ohne die Berücksichtigung der Erdkrümmung abgebildet werden. Da die Erde ein mehr oder weniger kugelförmiges Gebilde[39] ist, verzerrt eine Darstellung globaler Netze immer dann, wenn man den Schnittpunkt festlegen muss. Eine Weltkarte stellt an ihren Rändern einen Abstand her, der in Wirklichkeit nicht existiert. So zieht eine Projektion, die den Greenwich-Standard Median als Zentrum nimmt, den amerikanischen und den asiatischen Kontinent maximal weit auseinander. Das erzeugt eine visuelle Distanz, die rein geografisch nicht existiert. Eine Kante zwischen Tokio und Hawaii müsste dabei über die ganze Karte gehen, obwohl sie in einer anderen Projektion nur kurz wäre. Des Weiteren würde diese Kante viele andere Kanten kreuzen.
Bei einer rein geografischen Darstellung kann es zum Problem der Überdeckung kommen. Zentren, in denen viele Verbindungen bestehen, sind nicht wirklich gut zu überblicken da sich viele informationen auf kleinem Raum ballen. Vereinzelte Knoten oder Orte sind dagegen überschaubar. In der Praxis liegen deshalb Darstellungen von Bahnnetzen meistens auch in einer Form vor, die nicht exakt geografisch ist. Dies hat den Effekt, dass die Peripherie gestaucht und das Zentrum gestreckt wird. Der Vorteil dieser Darstellung liegt darin, dass die Kartengröße insgesamt abnimmt und die Ordnung der Stationen dadurch nicht verändert wird. Das ist vergleichbar mit einer logarithmischen Darstellung, wie sie beispielsweise in Funktions- und Scatter-Plots verwendet wird.
Die Verwendung von geografischen Distanzen ist in zwei dimensionalen Abbildungen niemals vollkommen unproblematisch. Die Erdkugel wird auf eine Fläche projiziert und die dazu verwendete Projektion verzerrt die geografische Position. Seit der Antike besteht die Erkenntnis, dass die Erde eine Kugel ist und die genaue Bestimmungen von Längen und Breitengraden anhand der Sterne war möglich.[83, 257] Trotz dieser genauen Methodik lassen sich die damals entstandenen Karten heute nur schwer lesen[40], da sich aufgrund der Sozialisation eine klare Lesart von Karten gebildet hat. Hierzu gehört die Ausrichtung nach Norden sowie die Wahl des Äquators und des Greenwich Medians als Zentrum einer Karte. Die Fixierung auf den europäischen Raum im Zentrum der Karte ist jedoch nicht erforderlich. Amerikanische Institutionen verwenden beispielsweise in Symbolen[41] Darstellungen mit dem amerikanischen Kontinent in der Mitte der Welt-Karte. Eine Erklärung findet sich in der strategischen Position, welche aus einer Karte mit anderem Zentrum besser hervor geht. In Abschnitt 2.2.4 wurde dies anhand des Risiko-Spielbretts diskutiert. Die UN verwendet verwendet im Gegensatz dazu eine Projektion der Erde mit dem Nordpol als Mittelpunkt um eine eine vereinnahmung der Mitte zu umgehen.[42] Wie sich das wirklich geografische Modell der Welt von der geistig "strukturierten" Welt unterscheiden

[39] Die Erde ist keine perfekte Kugel. Ein Geoid stellt dabei die mathematische Annäherung an die Form der Erde dar. Die Abweichungen, die sich daraus ergeben, werden im Weiteren übergangen.

[40] Die Karten von Claudius Ptolemaeus enthalten u.a. starke Verzerrungen.[257]

[41] Als Beispiele dieser Darstellung können die Embleme des United States Space Command und das Siegel des United States Fleet Forces Command `http://www.public.navy.mil/usff/Pages/logo.aspx`(zuletzt gesehen 20.03.2017) genannt werden. Sie beinhalten Weltkarten mit dem Amerikanischen Kontinent im Zentrum.

[42] `http://www.un.org/en/sections/about-un/un-logo-and-flag/index.html` zuletzt gesehen 20.03.2017.

kann. Mittelalterliche und antike Karten mussten sich nicht mit dem Fakt befassen, dass die Erde wirklich eine Kugel ist. Es hatte noch keine Erdumrundung stattgefunden, dadurch entfiel das Problem des Kartenrands, wie mit dem vollständigen Wissen um die Welt auftritt. Diese frühen Weltkarten sind daher im heutigen Verständnis eher lokale Karten. Im weiteren werden die "Ebstorfer Weltkarte" und an der "Tabula Peutingeriana"[262](Abbildung 2.11) vorgestellt.
Die "Ebstorfer Weltkarte" ist wahrscheinlich im 13. Jahrhunder nach Christus entstanden.[149, 122] Auf ihr sind nicht nur rein geografische Informationen enthalten, sondern auch geschichtliche Sachverhalte. Sie zeigt die Welt mit dem Kopf und den Gliedmaßen von Jesus Christus am Rand der runden Karte. Im Zentrum wird Jerusalem dargestellt. Dies lässt auf eine politische Zentrierung der Karten schließen. Auch die Anordnung der Orte zueinander ist nicht unabhängig von ihrer topologisch-geografischen Anordnung, jedoch so stark verzerrt das der Leser einige mühen aufwenden muss sie mit seinem heutigen geografischen Verständnis ins Verhältnis zu setzen.
Die "Tabula Peutingeriana" entstand wahrscheinlich auch im 13. Jahrhundert nach Christus, das zugrunde liegende Kartenmaterial ist jedoch sehr viel älter und wird auf das vierte Jahrhundert nach Christus geschätzt[122].[122, 262] Die Tabula Peutingeriana ist eine Karte, die Handelsrouten und Pilgerpfade beschreibt.[122] Diese Art der Karten ist distanzbasiert und anscheinend für den praktischen Gebrauch gemacht.[13, s. 37 f] Die "Tabula Peutingeriana" baut auf einer Menge von Einzelbeobachtungen auf. Die Positionen auf dieser Karte sind auf Grundlage einer erfahrenen Reisedistanz sehr wahrscheinlich adäquat, auch wenn sie nicht der geografischen Distanz entsprechen. Das Gesamtbild der "Tabula Peutingeriana" Welt entsteht aus einem abstrakten Zusammenspiel der Einzeldistanzen, die eigentlich als Listen geführt wurden. Der Aufbau von Graphen und darauf angewandte kräftebasierte Positionierungsverfahren nutzen ähnliche Verfahren. Multiple lokale Distanzen, Einzelbeobachtungen werden zur Schaffung eines Gesamtbildes verwendet.

2.4.3 Kräftebasierte Positionierungsverfahren

Kräftebasierte Positionierungsverfahren[43] stellen eine Möglichkeit dar, Knoten und Kanten automatisch zu positionieren. Hierbei handelt es sich um ein Modell aus anziehenden und abstoßenden Kräften die eine Ordnung herstellen.
In diesem Verfahren geben die Verbindungen vor, wie Knoten positioniert werden. Knoten ziehen sich an und stoßen sich ab. Dieses Vorgehen entspricht der Simulation von Anziehungskräften durch mechanische Federn und Abstoßungskräften durch magnetische Felder. Die Abstoßung aller Knoten untereinander wird dabei von den Anziehungskräften unterbunden. Diese Anziehungskräfte werden ins Modell einbezogen wenn zwei Knoten durch eine Kante verbunden sind. Bei diesem Vorgehen handelt es sich nicht um einen speziellen Algorithmus, sondern um eine Klasse von Algorithmen.[99, 132, 259]

[43] Engl.: "spring embedder", "forcedirected layouts".

Der Algorithmus läuft iterativ ab. Es wird schrittweise eine physikalische Simulation durchgeführt. Es gibt dabei nur zwei Verarbeitungsschritte.

1. Die Kräfte, die auf die Knoten einwirken, werden berechnet.
2. Die Knoten werden nach den in Schritt 1 berechneten Kräften verschoben.

Eine frühe Implementation dieses Vorgehens ist der Fruchtermann-Reingold Algorithmus. In diesem Modell stoßen sich alle Knoten gegenseitig ab. Die Kanten werden wie mechanische Federn simuliert, woher auch der Name Spring Embedder Algorithmus stammt. Auf der einen Seite bewirken die Federn, dass sich verbundene Knoten anziehen. Dabei wird die Anziehungskraft umso stärker, umso mehr die Feder gespannt ist. Zudem können die Kanten nicht beliebig gestaucht werden. Auf diese Weise wird ein Mindestabstand zwischen den Knoten hergestellt.[99]
Diese Form der Knoten-Positionierung ist dabei nie wirklich beendet. Es ist also kein wirklich deterministischer Algorithmus. Es gibt jedoch Abbruchbedingungen, die das Feststehen eines Systems in einem stabilen Zustand registrieren. Dies kann anhand der Energie, die im System herrscht, errechnet werden. Das tritt ein, wenn die Kräfte, die im ersten Schritt berechnet werden sehr gering sind und die Knoten nicht mehr stark verschoben werden. Die Positionsänderung der Knoten in einem Simulationsschritt ist dann nur noch marginal.
Die Verwendung physikalischer Systeme hat den Vorteil, dass die Gewichte von Kanten nativ mit in die Systeme übernommen werden können. Sie steuern dabei die Anziehungskräfte und Minimaldistanzen der Federn. Hier unterscheiden sich die Ansätze verschiedener Algorithmen. Dabei kommen die Grundannahmen von guter Positionierung ins Spiel, welche den einzelnen Algorithmen zugrunde liegen.[243, 388 f.]
Zu beachten ist dabei, dass diese Art von Algorithmen nicht die Ausgangsposition der Knoten festlegen. Daher muss man eine Anfangsanordnung finden, bevor die Knoten vom Algorithmus positioniert werden können. Das hat Auswirkungen auf das Endergebnis, da eine an sich schon stark verstrickte Ausgangsstruktur nicht zwangsläufig durch den Algorithmus geordnet wird.
Ein weiterer Vorteil solcher Verfahren ist die Kontinuität der Visualisierung, dadurch sind dynamische Visualisierung möglich.[71, 41] Auf diese Weise können Knoten und Kanten entfernt, verändert und hinzugefügt werden. Der Algorithmus reagiert direkt auf diese Eingriffe. Bei vielen interaktiven Graphenvisualisierungen kann der Nutzer auch in die Positionierung eingreifen. Dies geschieht bei den meistens durch das Anwählen und Ziehen von Knoten.[44] Der Nutzer kann so Überlagerungen, die vom Algorithmus nicht aufgelöst werden, selbst auseinanderziehen.
Die schrittweise Visualisierung ist für die Darstellung von Zeitgraphen wichtig. Das schrittweise Verändern der Knotenanordnung hält die Orientierung des Nutzers aufrecht. Würde nach einem Schritt eine komplett neue, optimale Ausrichtung aller Knoten erfolgen, müsste der Nutzer sich wieder vollkommen neu orientieren. Diese Systeme werden in der Prozessvisualisierung in der Softwareentwicklung verwendet.[57, 52]

[44] Engl.: drag and drop

Über die Jahre wurden eine Menge dieser Algorithmen entwickelt und in der praktischen Implementation stehen oft eine Menge verschiedener Features zur Positionierung bereit. Aufgrund der Menge der Features und des nicht deterministischen Verfahrens sowie der Ausgangspositionierung wird die Einstellung dieser Algorithmen immer auch ein Stück praktischer Optimierung bleiben. Die genauen Einstellungen bleiben abhängig von den zu verarbeitenden Daten. Hier sind Faktoren wie die Menge der Knoten oder die Menge der Kanten und deren Gewicht besonders zu beachten.
Grenzen hat das Verfahren häufig bei großen Graphenstrukturen. Hier stoßen die meisten Algorithmen aufgrund des großen Speicherbedarfs und Rechenaufwands an eine praktische Grenze. Für diesen Zweck gibt es Algorithmen, die auf große Graphen spezialisiert sind, wie beispielsweise Openord[161]. Bei dichten und großen Graphen gibt es das Problem der "Hairballs". Dies sind Visualisierungen, die nicht mehr sinnvoll zu lesen und zu positionieren sind und daher wie Haarklumpen aussehen. An dieser Stelle muss ein sinnvolles Relevanz- und Filterkriterium gefunden werden.[185]

Überlagerungen

Die Positionierung und Darstellung von Kanten und Knoten kann zum Problem für die Lesbarkeit von Visualisierungen werden. Damit die Darstellung eines Graphen gelesen werden kann, ist es wichtig, Knoten und Kanten zu erkennen und auseinanderhalten zu können. Bei Knoten ist das Problem noch recht einfach zu lösen. Hier werden sich überdeckende Knoten auseinander geschoben. Ein weiteres Problem stellt das Überlagern der Knotenbeschriftung dar.
Die Überschneidung von Kanten ist unter diesem Gesichtspunkt problematischer. Vor allem wenn viele Kanten in steilem Winkel übereinander laufen, sind diese nur noch schwer zu unterscheiden.
Hierfür gibt es leider keine triviale Lösung und eine vollkommen überschneidungsfreie Darstellung ist nicht immer machbar. Ein Graph, der ohne Kantenüberschneidungen dargestellt werden kann, wird planar genannt. Die Optimierung dieses Problems ist dabei nicht trivial und zwischen dem algorithmischen Erkennen von Überschneidungen und der perfekten Anordnung zur maximalen Überschneidungsfreiheit besteht ein Unterschied.
Selbst für den Menschen ist es, bei einer überschaubaren Menge von Knoten und Kanten schwer, überhaupt eine optimale Lösung zu finden. In Bezug auf die kräftebasierten Algorithmen können die Anfangsanordnungen dazu beitragen von Anfang an Kantenüberlagerungen zu vermeiden.
Im Design von Schaltkreisen kommt es zu ähnlichen Problemen. Dort hat das Überlagern von Leitungen nicht nur visuelle Konsequenzen sondern auch funktionale. Auf einer Platine lassen sich Überschneidungen nur in begrenztem Maß realisieren. Dies ist etwa durch mehrere Schichten von Leiterbahnen in einer Platine oder eine Drahtbrücke möglich. Für die Abbildung solcher Fälle gibt es eigene Zeichen-Konventionen in der Symbolsprache des Schaltkreisdesigns.

3D

Dreidimensionale Darstellungen von Graphen sind problemlos möglich.
Iterative physikalische Modelle funktionieren auch in dreidimensionalen Räumen. Die Kräfte ziehen und drücken dann in eine weitere Dimension.[45]
Ein vermeintlicher Vorteil könnte hier darin gesehen werden, dass sich Kanten in einem dreidimensionalen Raum nur mit geringer Wahrscheinlichkeit kreuzen. In einem zweidimensionalen Raum ist es dagegen wahrscheinlich, dass sich zwei Linien kreuzen.
Das Problem, einer dreidimensionalen Darstellungen liegt darin, dass es für den Menschen immer nur eine Perspektive oder Perspektiven auf die dreidimensionalen Modelle gibt. Jeder dieser Perspektiven oder Blickwinkel wäre immer mit einer großen Menge von verdeckten Informationen verbunden. Im Hinblick auf die Überschneidungen wäre es daher noch komplexer, da ein Minimum an Überschneidungen für jede Perspektive errechnet werden müsste.
Die in der räumlich explorativen Erfahrung sind jedoch ein Mittel Sachverhalte besser zu erfassen. Die kontinuierliche räumliche Erfahrung ist dabei geeigneter, als den Raum durch Einzelbilder zu erschließen. Dies gilt besonders für die Beobachtung von Strukturen aus einer Vogelperspektive wie in den Theorien zur Wahrnehmung von James J. Gibson beschrieben.[105, 108] Objekte können vom Betrachter durch die Bewegung und Veränderung in Größenverhältnisse gesetzt werden. Daher wäre anzunehmen, dass die dritte Dimension der Natur durch die Nähe zur Welt, in welcher der Mensch lebt, einen Mehrwert bringen könnte.
Nach Piaget entsteht die euklidische Raumvorstellung aus der frühkindlichen Erfahrung des Menschen mit seiner Umwelt. Es muss aber auch beachtet werden das laut Piaget die Erkenntnis des euklidischen Raums aus einer Einschränkung des topologischen Raums hervor geht.[197]
Daten und auch Relationen aus dem topologischen Raum lassen sich nicht immer sinnvoll im euklidischen Raum abbilden.
Ein Molekularmodell kann in drei Dimensionen abgebildet werden, da seine Teile wirklich im Raum angeordnet sind. Selbst wenn nur Relationen, aber keine Positionen bekannt wären, dann müssen alle Teile eine Position im euklidischen raum Einnehmen. Hier könnte wahrscheinlich ein rein kräftebasierter Algorithmus zu ordentlichen Ergebnissen führen.
Bei einer Abbildung von sozialen Beziehungen oder wissenschaftlichen Zitaten, Ähnlichkeiten und Verwandtschaftsverhältnissen etc. ist das nicht der Fall. Diese Verhältnisse und Entitäten die sie Beschreiben existieren im Euklidischen raum. Doch sind ihre konzeptionellen Verbindungen nicht an ihre Repräsentanten gebunden. Werden solche Beziehungen aus dem topologischen Raum im dreidimensionalen, euklidischen Raum abgebildet, dann entsteht keine Anordnung, die mit den erlernten Vorstellungen von Raum übereinstimmen müssen.
Teilverstöße gegen die Regeln des euklidischen Raums mögen zwar teilweise abbildbar sein,

[45] Das Einführen einer vierten und fünften Dimension wäre rein mathematisch auch möglich. Allerdings bestünde dann das Problem, dass in diesem Fall nicht mehr von einer Visualisierung gesprochen werden kann. Dieses Konstrukt wäre dann höchstens eine Anordnung im n-dimensionalen Raum.

doch ließ sich ein solcher Raum nur schwer "Erleben".[46] Im schlimmsten Fall würde der Vergleich der dreidimensionalen Projektion mit dem echten Raum zu Fehlschlüssen durch die real-weltliche "Erfahrung" führen.
Es ist nicht auszuschließen, dass es wirklich sinnvolle dreidimensionale Anordnungen geben kann. Es ist aber mehr als fraglich, ob ein simpler kräftebasierter Ansatz für die Anordnung im Raum ausreichend ist. Die Anpassung der Raumnutzung könnte das ändern. Die dritte Dimension könnte beispielsweise nicht frei für den kräftebasierten Algorithmus positionierbar sein sondern, wie bei einem Scatterplot, eine Eigenschaft eines Knotens abbilden. Diese Modelle würden über ein einfaches Übertragen eines kräftebasierten Algorithmus in den dreidimensionalen Raum hinaus gehen.

2.4.4 Bäume

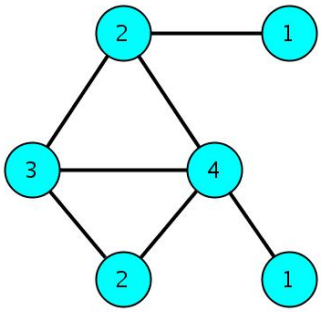
Bäume sind Sonderformen von Graphen. Sie lassen sich algorithmisch einfacher und reproduzierbarer abbilden und anordnen als vollständige Graphen mit Kreisen. Kann ein Graph in einen Baum umgewandelt werden, dann lässt er sich einfacher darstellen.[80] Durch ihre eingeschränkten strukturellen Merkmale gibt es eine Vielzahl von Algorithmen, die Bäume effizient und überschneidungsfrei anordnen können.
Bäume lassen sich aber nicht nur als Netzdiagramme darstellen. Treemaps[229] basieren nicht auf Darstellungen mit Knoten und Kanten, sondern auf ineinander angeordneten Kästen. Eine Spielart dieser Darstellungsform sind Treemaps, die auf Voronoi Diagrammen beruhen. So können organisch wirkende Treemaps erstellt werden.[14]
Die Interaktive Ansichten von Treemaps können im Weiteren durch Zoomen fokussiert werden. Dabei können Untereinheiten, die in der Anfangsansicht vernachlässigt wurden, dynamisch nachvisualisiert werden.

2.5 Bewertung und Merkmalsbeschreibung

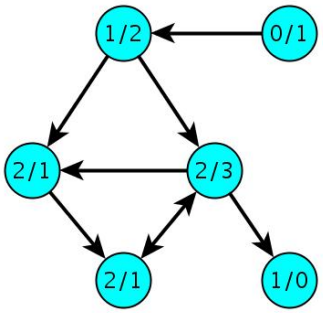
Wie gezeigt, können Modelle von Graphen verschiedene Ausprägungen haben. Für die Bewertung von existierenden Graphen, die diese Merkmale und andere Parametern aufweisen, werden gängige Mittel gebraucht, um die Eigenschaften von Graphen zu beschreiben. Hier werden nun verschiedene Modelle zur Bewertung der strukturellen Eigenschaften von Netzen gezeigt.
Diese Maßzahlen und Wertungen sind nicht in jedem Fall sinnvoll einzusetzen. Auf die Durchdringung der abstrakten Aussage eines errechneten Wertes muss auch immer eine Interpretation der konkreten Daten folgen. Je nach Anwendung und Daten sollten im Vorhinein die Verfahren und Maßzahlen, welche verwendet werden, sorgfältig ausgewählt werden.

[46] Dieser Ansatz wird in der Kunst bei Escher verwendet, um faszinierend verwirrende Werke zu schaffen. Bei H.P. Lovecraft wird der "nicht euklidische Raum" als fantastisches Schreckenselement des Fremden verwendet.[158, 159]

Abbildung 2.12: *Beispiele für Gradzahlen von Knoten*



(a) *Ungerichteter Graph mit Gradzahlen auf Knoten*



(b) *Gerichteter Graph mit Gradzahlen auf Knoten. Die erste Zahl im Knoten gibt den eingehenden Grad und die zweite Zahl den ausgehenden Grad des Knotens an.*

Des Weiteren können die Merkmale und Bewertungen unterschiedlich schnell berechnet werden. Dabei ist es wichtig zu beachten, zu welchem Zweck ein Maß berechnet werden soll. Ein simples Maß wie die Dichte oder die Gradzahl kann auch für große Netze schnell berechnet werden. Bei komplexeren Maßzahlen, wie den Zentralitätsmaßen nimmt die benötigte Rechenzeit mit der Größe und Dichte des Graphen stark zu.

2.5.1 Grad

Theoretisch wie praktisch ist der Grad eines Knotens das simpelste Maß eines Graphen. Der Grad[47] wird durch die Menge der an einen Knoten anhängenden Kanten berechnet. Bei ungerichteten Kanten gibt es einen Grad, die Anzahl der Kanten, an die ein Knoten angeschlossen ist. Knoten ohne Anbindung werden als isoliert bezeichnet. In Abbildung 2.12a ist dies beispielhaft abgebildet.
Wird ein gerichteter Graph betrachtet, gibt es drei Gradzahlen. Dies sind die Menge der eingehenden, der ausgehenden und die Menge aller Kanten. In Abbildung 2.12b ist beispielhaft angeführt, wie sich die Ein- und Ausgangsgradzahlen in gerichteten Kanten verteilen. Der Gesamtgrad aller Kanten kann als Summe der eingehenden und ausgehenden Kanten beschrieben werden.

2.5.2 Dichte

Die Dichte[48] eines Graphen beschreibt das Verhältnis zwischen der Anzahl der vorhandenen Kanten und der maximal theoretisch möglichen Anzahl von Kanten. Die Dichte besagt, wie viele der maximal möglichen Kanten im Graphen vorhanden sind. Dieses Maximum ist bei den verschiedenen Arten von Netzen unterschiedlich zu berechnen.
Solange keine parallelen Kanten von verschiedenen Typen im Modell erlaubt sind, verhält

[47] Engl.: degree

[48] Engl.: density

sich die Menge der Kanten generell quadratisch zur Menge der Knoten. Die Richtung der Kanten spielt hier im Weiteren eine wichtige Rolle. Ein gerichteter Graph kann doppelt so viele Kanten haben wie ein ungerichteter Graph mit der gleichen Anzahl von Knoten. Wieder anders errechnet er sich nach der hier genutzten Definition für bi- oder multipartiten Netze, da Kanten nur zwischen den Kategorien existieren können. Eine Menge der ansonsten möglichen Verbindungen ist so per Definition ausgeschlossen. Im Fall eines bipartiten Netzes ist das die Menge der Knoten des ersten Typs multipliziert mit der Menge der Knoten des zweiten Typs.[36]
Da die Dichte informatisch schnell zu berechnen ist, wird sie gerne als Grundinformation über einen Graphen verwendet. In diesem Zusammenhang wird auch von stark- und dünnbesetzten Graphen gesprochen. Es kann aber allgemein gesagt werden, dass Netze mit einer hohen Dichte, also stark besetzte Graphen, eher als uninteressant angesehen werden können. Hier erschließt sich kaum Information aus einer Netzstruktur, da jeder Knoten mit jedem anderen verbunden ist.

2.5.3 Komponenten

Ein weiteres der simplen Merkmale ist die Komponentenanzahl. Als Komponente wird ein unverbundener Teilgraph betrachtet. Definiert werden kann dies als Menge von Knoten, welche nur untereinander verbunden sind. Knoten in einer Komponente können nicht von Knoten außerhalb der Komponente erreicht werden.
In vielen Analysen wird die größte Komponente[49] alleinig für die Analyse genutzt. Dies ist oft nur ein kleiner Teil der Gesamtdatenmenge. Der Hintergrund dieses Vorgehens kann in einer Abgrenzung aufgrund der Datenlage gesehen werden. Außerdem sind manche Maßzahlen bei einem Graphen aus mehreren Komponenten nicht vergleichbar, sie variieren nach Komponentengröße.[50]
Eine weitere Komponenteneinteilung finden die Teilgraphen in der Definition von stark verbundenen Komponenten. In gerichteten Graphen sind das Komponenten, die sich nicht nur durch einfache Verbundenheit auszeichnen, sondern auch der Kantenrichtung folgend verbunden sind.
Dem entgegen steht die schwach verbundene Komponente. Dies ist eine Komponente, in der alle Knoten untereinander erreichbar sind, ohne dass die Richtungen der Kanten berücksichtigt werden.

2.5.4 Zentralitäten

Zentralitätsmaße versuchen, die Wichtigkeit und Verbundenheit von Knoten darzustellen. Dabei soll anders als beim Grad die Gesamtstruktur des Graphen in die Bewertung einbezogen werden. Die wichtigsten Zentralitätsmaße, Betweenness Centrality und Closeness Centrality basieren auf Überlegungen von Kommunikationsnetzen. Im Gegensatz zum Grad wird mehr Aufwand betrieben, um eine nicht lokale Bewertung zu erlangen.[96]

[49] engl. giant component

[50] Beispielsweise die Closeness Centrality, wie sie in Abschnitt 2.5.4 beschrieben wird.

Zentralitätsmaße schaffen eine lineare Gewichtung von Knoten aus der eigentlich nicht linearen Anordnung von Graphen. Dabei streuen die Maße unterschiedlich gut, es kann daher auch als ein Ziel eines Zentralitätsmaßes gesehen werden, wie gut die Bewertungen für jeden einzelnen Knoten und jede einzelne Kante unterscheidbar sind. Es geht dabei um die Rangfolge, die aus einem Zentralitätsmaß hervorgeht. Es werden, anders als beim Grad, wenige Knoten oder Kanten mit der gleichen Zahl bewertet.

Betweenness Centrality

Ein Betweenness Centrality Wert zeigt die Möglichkeit der Kommunikationskontrolle eines Knotens an. Ein Knoten mit einer hohen Betweenness Centrality hat eine hohe Wahrscheinlichkeit dafür, dass eine Nachricht zwischen zwei beliebigen anderen Knoten über ihn läuft.
Mathematisch bedeutet dies, dass der Knoten auf einer großen Menge von kürzesten Pfaden zwischen allen anderen Knoten liegt. Der Parameter kommt aus dem Umfeld der Soziologie und wurde von Freeman eingeführt.[96]
Freeman beruft sich dabei auf Bavelas.[22] Wenn es wahrscheinlich ist, dass eine Nachricht über einen Knoten läuft, dann hat dieser Knoten auch großen Einfluss darauf, die Kommunikation zu kontrollieren.
Mathematisch wird dies mit dem Vorkommen des Knotens auf den kürzesten Pfaden zwischen allen anderen Knotenpaaren argumentiert.
Der Knoten kann die Nachricht, die über ihn läuft, beeinflussen. Er kann beispielsweise durch das Blockieren und Verändern der Information die Kommunikation kontrollieren.
Hier ist zu beachten, dass jeder Pfad als gleich wahrscheinlich angenommen wird. Dabei wird ignoriert, dass es feste Routen oder Ähnliches gibt. Das kann jedoch je nach Berechnung der Faktoren und Gewichte beeinflusst werden.
In ihrer Eigenschaft als Bewertungskriterium lässt sich die Betweeness Centrality gut zur abstufenden Bewertung der Knoten benutzen. Es wird eine hohe Merkmalsausprägung angenommen, das heißt, in der Bewertung gibt es wenige Knoten, die mit der gleichen Zahl bewertet werden.
Die Ausprägung dieses Wertes ist in sich selbst nur komparativ im Graphen und nicht absolut interpretierbar. Eine Normierung des Faktors zwischen 0 und 1 ist jedoch möglich.
Die Berechnung der Betweeness Centrality ist eher aufwendig. Der Algorithmus von Brandes für ungerichtete Graphen hat eine Laufzeit, die sich nach Knotenanzahl mal Kantenanzahl richtet.[38]

Closeness Centrality

Diese Form der Zentralität wurde auch von Freeman im gleichen Atemzug wie die Betweenness Centraltiy formuliert.[96]
Die Closeness Centrality ist eigentlich die inverse Eigenschaft der “Farness”. Dieser “Entfernungswert” ergibt sich aus der Summe aller kürzesten Pfade zu allen anderen Knoten im Netz. Invertiert man diesen Wert, dann bekommt man die Closeness Centrality.
Freeman beschreibt die Eigenschaft eines Knotens mit hoher Closeness Centrality als un-

abhängig. Es können alle Nachrichten über kurze Wege zum Knoten gelangen, ohne dass die Nachrichten über viele andere Knoten laufen müssen.
Im Gegensatz zur Betweenness Centrality sind die Werte, die dieses Zentralitätsmaß annehmen kann, geringer. Es werden also eine Menge von Knoten mit dem numerisch gleichen Wert bewertet.
Des Weiteren lässt sich die Closeness Centrality immer nur für eine Komponente errechnen. Wenn ein Knoten nicht in einer Komponente enthalten ist, kann er nicht erreicht werden, insofern ist er unendlich weit entfernt. Die meisten Algorithmen errechnen daher eine separate Closeness Centrality für jede Teilkomponente. Eine kleinere Komponente kann dabei im Durchschnitt eine bessere Erreichbarkeit haben, obwohl die meisten Knoten gar nicht erreichbar sind. Die Closeness Centrality eignet sich nicht als komparatives Maß für Knoten, die nicht in der gleichen Komponente liegen. Das wird in Visualisierungen sichtbar, denn kleinere Komponenten werden als "signifikanter" dargestellt.

2.5.5 Clustering

Als Cluster sind Mengen von Knoten zu verstehen. Cluster werden anhand von Eigenschaften des Graphenmodells zusammengefasst, beispielsweise aufgrund ihrer starken Verbundenheit untereinander. Die automatische Extraktion solcher Cluster kann dabei in verschiedenen Methoden erreicht werden. Im Gegensatz zu Komponenten lassen sich Cluster in verbundenen Graphen bestimmen.
Cluster oder auch Communities innerhalb von Graphen werden anhand ihrer Verbindungsstruktur und der lokalen Dichte ihres Umfelds bestimmt. Alternativ können Cluster aber auch anhand von nichtrelationalen Eigenschaften gefunden werden.[266, Abschnitt 6.8]

Nichtrelationales Clustering

Es gibt eine Reihe von Clusteringverfahren, die aufgrund ihrer Distanz in einem Raum, der aus zwei oder mehr Eigenschaften gebildet wird, hervorgehen.
Hier sind zentrumsbasierte Clusteringverfahren zu nennen. Für diese Verfahren werden keine Kanten benötigt, sondern nur Knoten (Punkte), die eine Position in einem Raum oder auf einer Fläche besetzen. Anhand der Entfernung eines Knotens zu einem gewählten Zentrum wird bestimmt, in welches Cluster ein Knoten gesteckt wird.
Eine relationale Methode die daraus hervorgeht basiert auf Ähnlichkeitsnetzen. Dafür wird eine Ähnlichkeitsfunktion aufgestellt, auf deren Grundlage Knoten die Knoten verbunden werden. Die Ähnlichkeitsfunktion gibt vor, wie ähnlich zwei Werte sein müssen, damit sie eine Verbindung erhalten. Am Ende können so die verbundenen Komponenten als Gruppen definiert werden. Ein beliebter Algorithmus, der diese Eigenschaften bedient, ist DBSCAN.[89]

Hierarchisches Clustering

Im hierarchischen Clustering werden die einzelnen Knoten aufgrund eines Ähnlichkeitsmaßes zusammengefasst. Das Ähnlichkeitsmaß wird durch die Kanten beschrieben. Hierarchische Clusteringverfahren bilden Bäume von Mengen.

Im hierarchischen Clustering gibt es einen aussagelosen Anfangs- und einen aussagelosen Endzustand. Alle Knoten stehen für sich oder alle Knoten sind einem Cluster zugeordnet. Dabei kommt es auf das Verfahren an, wie dieser Zustand aussieht. Entweder werden am Anfang alle Knoten einzeln gesehen und dann nach und nach zusammengefasst oder am Anfang sind alle Knoten in einem Cluster und dieses wird auseinanderdividiert.
Im ersten Fall ist der Ausgangszustand eine Menge von Clustern, die der Knotenanzahl entspricht. Diese wird dann immer weiter zusammengefasst, bis alle Knoten im gleichen Cluster sind. Dazwischen bilden sich verschiedene Clustermengen.[94]
Die interessanten Clustermengen zu finden, stellt bei hierarchischen Clusteringverfahren eine Herausforderung dar. Es muss in etwa bekannt sein, wie viele Cluster in der Ergebnismenge erwartet werden. Dann kann der Algorithmus bei Annäherung an das Ergebnis angehalten werden.
Ein Algorithmus zur Erkennung von Communities, der hierarchisch arbeitet, ist der Fast Unfolding of Communities-Algorithmus.[34] Sein Ansatz basiert auf der Maßzahl für die Güte von Clustern, wie sie von Newman vorgestellt wurde.[180] Die Maßzahl von Newman zeigt an, wie stark die Knoten in verschiedenen Clustern untereinander verbunden sind. Am Anfang sind alle Knoten einzelnen in ihrer eigenen Community.
Im ersten Schritt werden nach und nach die einzelnen Knoten mit den Communities ihrer Nachbarn zusammengefügt. Dabei wird geschaut, welche Zusammenführung die Güte nach Newman maximiert.
Im zweiten Schritt werden die gefundenen Communities zu Metaknoten zusammengefasst, deren Selbstkante die Menge der Kanten in der Community repräsentiert. Anschließend werden diese Metaknoten wieder in Schritt 1 bearbeitet.[34]
In jeder Iteration werden es weniger Communities, da diese immer weiter zusammengefasst werden. Am Ende sind alle Knoten in der gleichen Community.
Bei diesem Algorithmus hat die Ordnung und damit die Abarbeitungsreihenfolge der Knoten Einfluss auf die entstehenden Communities.

Clusterkoeffizient

Der Clusterkoeffizient[51] bildet keine Cluster. Er bewertet vielmehr, wie stark ein Netz zu Clustern tendiert. Er ist ein globales und lokales Maß für einen Graphen. Ein Clusterkoeffizient kann also für einen ganzen Graphen und für einen einzelnen Knoten errechnet werden. Clustering bezieht sich auf die kleinsten Kreise, die in einem Graphen auftauchen — die Dreiecke.
Ein Knoten hat einen hohen lokalen Clusterkoeffizient, wenn viele seiner Nachbarn auch untereinander verbunden sind. Es handelt sich dabei um den Anteil der Nachbarn, die auch mit anderen Nachbarn verbunden sind.[260]
Knoten mit wenigen Nachbarn können einen hohen Clusterkoeffizient erreichen, während Knoten mit vielen Nachbarn nur schwer hohe Werte erreichen können, da es viele potenzielle Verbindungen gibt. Hier spielt der Grad eines Knotens eine wichtige Rolle, aber auch,

[51] Engl.: clustering coefficient

wie dicht ein Graph ist. Mit der Dichte steigt die Wahrscheinlichkeit, dass sich Dreiecke bilden.
Bipartite Netzwerke weisen aufgrund ihrer Struktur keine Dreiecke auf. Daher haben diese Graphen auch immer einen Clusterkoeffizienten von Null. Diese Form von Netzen braucht für die Berechnung vergleichbarer Maßzahlen, spezielle Verfahren. Diese basieren nicht auf Dreiecken und weichen von Verfahren, wie sie bei One-Mode-Netzen verwendet werden, ab.[36]

2.5.6 Zufall und Wirklichkeit

Wenn es um die Konstruktion von Netzen geht, steht immer wieder die Frage im Raum, was ein echtes Netz ausmacht. Ist die Netzstruktur, die beobachtet wird, wirklich signifikant oder ist es eine Struktur, die auch zufällig entstehen könnte? Bei Zufallsnetzen steht dabei der Vergleich mit explizit erfassten Netzen im Mittelpunkt. Dabei können zufällige Graphen andere Verteilungen von Attributen aufweisen.

Konstruktion

Zufallsnetze werden im Allgemeinen aus einer vorgegebenen Menge von Knoten und Kanten konstruiert. Je nach Methode werden Kanten und Knoten schrittweise zum Graphen hinzugefügt. Dabei gibt es verschiedene Systeme, nach denen die Kanten hinzugefügt werden. Im Erdős–Rényi Model werden die Kanten zufällig hinzuzufügt.[209]
Zufällige Netze können auch so erstellt werden, dass sie die gleiche Verteilung in den Gradzahlen der Knoten aufweisen wie die erfassten Graphen. Das bietet die Möglichkeit zu sehen, ob eine Verteilung anderer Werte wie der Betweenness Centrality und der Closeness Centrality stark von zufälligen Verteilungen abweichen.
Durch das iterative Vorgehen bei der Erzeugung von zufälligen Graphen kann direkt eine Zeitdimension mit aufgezeichnet werden. Auf diese Weise lassen sich gerichtete Time Graphs erzeugen. So kann dynamisch verfolgt werden, wie sich beispielsweise die Dichte auf die Verbundenheit eines Graphen auswirkt.

Powerlaw und Skalenfreiheit

Das Powerlaw ist eine Aussage darüber, wie sich in Realweltnetzen die Verteilung der Links verhält. Dabei ist die genaue Aussage, dass der Grad von Knoten einer sogenannten Long-Tail-Verteilung entspricht. Es gibt wenige Knoten, die einen hohe Grad haben, den sogenannten Kopf[52], und viele Knoten, die wenige Verbindungen haben, den sogenannten Schwanz[53].
Um ein solches skalenfreies Netz[54] zu erzeugen, kann ein Preferential attachment Algorithmus verwendet werden.[16] Dabei werden Knoten, die viele Kanten haben, wahrscheinli-

[52] Englisch head

[53] engl. tail

[54] Engl.: scale free

cher Kanten zugewiesen als Knoten mit wenig Kanten. Des Weiteren können auch vorgegebene Verteilungen auf Zufallsnetze angewendet werden.
Diese Powerlaw-Verteilung von Daten verursacht dabei meist Probleme bei der Analyse von Netzen. Die großen Knoten verdecken die Aussagekraft der kleinen und weniger betrachteten Knoten.
Eine entsprechende Verteilung findet sich beispielsweise bei Links auf Webseiten. Im Falle von Webseiten werden beliebte Webseiten durch ihre Popularität und die dadurch entstehende Aufmerksamkeit bekannter. Das ist sozusagen ein selbstverstärkender Effekt.[15]

2.6 Abbildung von Graphen in der Informatik

Die vorgestellten Eigenschaften des Graphen-Modells und die vorgestellten Metriken müssen für die Verarbeitung im Computer abgebildet werden. Auf dem Computer, mit seiner linearen Verarbeitungsstruktur und der linearen Vergabe von Speicherzellen, lässt sich die nicht-lineare Struktur von Graphen nicht nativ abbilden. Die Datenstruktur hat Einfluss darauf, wieviel Speicher benötigt wird, was abgebildet werden kann, wie effizient Maßzahlen errechnet werden können und welche Algorithmen effizient auf ihr arbeiten können. Andere eigenschaften stellen die Effizenz von Änderungen am Graphen, also das Hinzufügen und das Entfernen von Knoten und Kanten, dar.

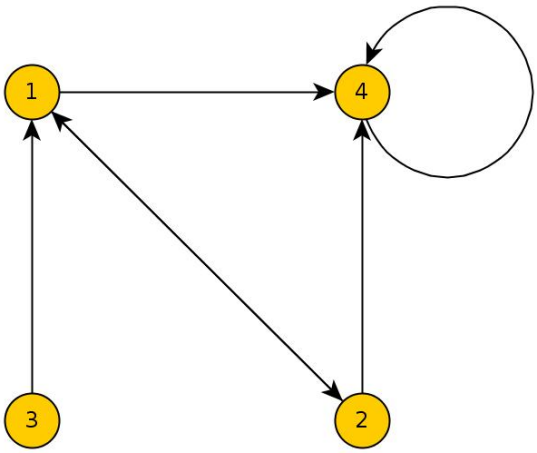
2.6.1 Speicherrepräsentationen

Computer müssen eine Modellierung auf Ebene des Arbeitsspeichers festlegen. Bei nichtlinearen Strukturen wie bei Graphen lassen sich verschiedene Speicherrepräsentationen vergleichen. Diese unterscheiden sich durch ihre Effizienz in der Verarbeitung und ihres Speicherverbrauchs. Bei nicht-linear sondern quadratisch verlaufenden Speicheranforderungen, können bei großen Graphen die Ressourcen knapp werden. Auch wenn angenommen werden kann, dass das Mooresche Gesetz[174] gilt und sich demnach die Rechen- und Speicherkapazitäten jedes Jahr verdoppeln, stellen quadratische Berechnungszeit und Speicherverbrauch ein nicht unerhebliches Problem dar.
Das Abbilden von Inhalten auf Graphen über die Struktur hinaus, wie Gewichte und Kategorien, aber auch Bilder und sonstige Informationen, ist meist unproblematisch. Im Arbeitsspeicher lassen sich jegliche Datenstrukturen mit direkten Speicherreferenzen anbinden. Sie verbrauchen auch Speicher, werden jedoch im Weiteren vernachlässigt, da sie sich nicht auf die Verarbeitungszeit der strukturellen Information auswirken.

Adjazenzmatrix

Eine Adjazenzmatrix (Abbildung 2.13b) ist eine Matrizendarstellung eines Graphen. Auf die Reihen und Spalten werden jeweils alle Knoten abgebildet. In den Zellen, die sich dazwischen abbilden, werden die Kanten markiert. Was genau in den Zellen steht, hängt von der Gewichtung der Kanten oder sonstigen Eigenschaften der Kanten ab.
Die Abbildung eines Graphen als Matrize hat eine lange Tradition. In der Mathematik beispielsweise wird die Errechnung von Maßzahlen und die Verarbeitung von Graphen

Abbildung 2.13



(a) *Visuelle Repräsentation des Graphen*

	1	2	3	4
1		X		X
2	X			X
3	X			
4				X

(b) *Darstellung des Graphen als Matrix*

Knoten	Nachbarn	
1	2	4
2	1	4
3	1	
4	4	

(c) *Darstellung des Graphen als Adjaziensliste*

Quelle	Ziel
1	2
1	4
2	1
2	4
3	1
4	4

(d) *Darstellung des Graphen als Liste*

meistens als Matrize angegeben. Auch das Blockmodelling basiert auf dieser Abbildungsform.
Matrizen lassen sich je nach Graphen-Modell verkleinern. Bei einem ungerichteten Netzwerk kann man den unteren Teil der Matrix weglassen, da es nur eine Verbindung zwischen den Knoten geben kann. In Abbildung 2.13b ist dieser Bereich grün dargestellt. In der Diagonalen der Matrix verlaufen bei diesem Verfahren die Selbstreferenzen. Diese können auch weggelassen werden, falls ein Modell keine Selbstreferenzen vorsieht. In Abbildung 2.13b ist dieser Bereich blau dargestellt.
Ein rein gerichteter Graph ohne Selbstreferenzen käme also mit weniger als der Hälfte der Matrix aus. In Abbildung 2.13b ist dieser Bereich gelb dargestellt.
Im Fall einer kompletten Darstellung eines Graphen mit allen hier beschriebenen Kanten verhält sich der Speicherverbrauch für einen Graphen quadratisch zu der Anzahl der Knoten in einem Graphen.[115, s.24] Hier ist eine direkte Verbindung zur berechnung der Dichte, da alle möglichen Kanten eine Zelle in der Matrix benötigen.
Die Darstellung einer Matrize im Rechner ist ein zweidimensionales Array.
Der Nachteil, der aus dieser Abbildung besteht darin, dass viele Graphen nur wenige Einträge in der Matrix besitzen. Hierfür gibt es Algorithmen und Datenstrukturen, die versuchen die Leerstellen in den Matrizen zu komprimieren. Diese Verfahren werden als "sparse-matrixVerfahren, also dünn besetzte Matrizen, bezeichnet.[59, Kapitel 22][117, s. 170ff]

Adjazenzliste

Eine Adjazenzliste (Abbildung 2.13c) speichert für jeden Knoten die anhängenden Knoten in einer Liste oder einem Array. Im Gegensatz zur Matrizendarstellung wird hier weniger Speicher verwendet. Er ist direkt abhängig von der Menge der Kanten. Dafür sind die Zugriffszeiten schlechter als bei der Matrix.[117, s. 172f]

Tupelliste

Die einfachste Methode der Speicherrepräsentation ist die Auflistung von Kanten als Liste von Tupeln (Abbildung 2.13d), die Start- und Endknoten beschreiben. Falls von Interesse, kann eine separate Auflistung von Knoten stattfinden.
Es werden dabei alle Kanten mindestens als Start und Ziel definiert. Die Reihenfolge, daher die Bedeutung von Start und Ziel, kann bei ungerichteten Kanten vernachlässigt werden.
Eine Liste von Knoten ist in diesem Fall nur dann wichtig, wenn sie weitere Informationen über die Knoten beinhaltet, wie beispielsweise Label, Gewicht oder Klasse. Die meisten Dateiformate zur Speicherung von Graphen lassen sich auf diese Repräsentationsform zurückführen.
Beim Durchlaufen des Graphen ist diese Darstellung ineffektiv, da für jeden Schritt alle Kanten durchlaufen werden müssen.

2.6.2 Datenrepräsentationen

Graphen können durch eine Vielzahl von Dateiformaten dargestellt werden. Dabei können mehr oder weniger Attribute betrachtet werden. Im Allgemeinen ist zu sagen, dass die Abbildung von Graphen auf der Dateiebene stark von den gewählten Eigenschaften des Graphenmodells abhängt,[55] die in der Datei abgelegt werden sollen.
Dabei ist zu beachten, dass ein XML-basierendes Format immer mehr Overhead hat als ein CSV basiertes Format[155]. Der Vorteil von XML liegt jedoch in der Geschwätzigkeit, die laut Design die Daten menschenlesbar und auch in Zukunft verständlich halten soll. Des Weiteren können in XML auch andere Metadaten problemlos hinterlegt werden. Sie werden im schlimmsten Fall von den verarbeitenden Programmen ignoriert.
Weiterhin wichtig für die Wahl des Formates ist das Anfügen von nicht zum Graphenmodell gehörenden Informationen zu Knoten und Kanten. Hier können Metadaten wie Zahlen oder Tupel im Sinne einer relationalen Datenbank gemeint sein oder sogar komplexere Strukturen wie Mengen und Listen.
Die hier angeführten Formate sind nicht erschöpfend. Es soll lediglich ein Überblick über die möglichen Abbildungen von Graphen und deren Ausdrucksstärke sowie Dokumentationsmöglichkeiten gegeben werden.

CSV/Textformate

In CSV Formaten[56] wird ein Graph oft durch zwei Dateien abgebildet. Die CSV-Datei für die Kanten und die CSV-Datei für die Knoten. Als Minimalanforderung kann man hier eine Liste von Kanten sehen. Diese würde sich nur aus zwei Spalten ergeben: dem Startknoten und dem Zielknoten. Wenn nur die Kanten abgebildet werden, dann ist über die Knoten nicht mehr Information im Modell als deren eindeutiger Identifier. Die Interpretation, ob es gerichtete Kanten oder ungerichtete Kanten sind, wird meist nicht explizit genannt.
Eine Vielzahl von Programmen definieren eigene Dateiformate, in denen mehr Informationen und einige Metainformationen gespeichert werden können, wie Pajek[204] und NWB[245] bzw. Sci^2[246]. Diese speziellen Dateiformate beinhalten jedoch meistens auch Definitionen, wie sie in den gängigen CSV-Repräsentationen verwendet werden.
Pajek nutzt ein verteiltes Modell, das Zusatzinformationen zu einem Graphen, wie etwa verschiedene Gewichte, auf mehrere Dateien verteilt. Dabei ist es dann wichtig, dass die Werte der Knoten und Kanten in der gleichen Reihenfolge bleiben, damit die Daten am Ende wieder richtig zusammengefasst werden können. Im Fall von Sci^2 und NWB, werden Knoten und Kanten in die gleiche Datei geschrieben und durch Steuerzeichen getrennt.

XML basierte Formate, GraphML

GraphML, die Graph Modelling Language, ist eine Auszeichnungssprache die auf XML beruht. Mit dieser Sprache können verschiedenste Modelle von Graphen abgelegt werden.

[55] Eine Beschreibung, der Eigenschaften von Netzwerkmodellen findet sich in Abschnitt 2.3.

[56] Engl.: “comma” oder “character seperated file”.

GraphML kann gerichtete sowie ungerichtete Kanten als auch einen Mix aus beiden abbilden. Auf Knoten und Kanten bezogene Attribute können aufgrund der erweiterbaren Struktur von XML in einer Vielzahl von Varianten vergeben werden. Spezielle Eigenschaften können durch eigene Tags oder der Verwendung eines speziellen Standards abgebildet werden.[39]
Ein Beispiel hierfür ist das Programm yEd, mit dem Diagramme erstellt werden können. yEd erweitert den GraphML-Standard um ein eigenes Darstellungsvokabular. So sind die Standardausgabedateien durch eigene Auszeichnungen beliebig erweiterbar, während gleichzeitig alle strukturellen Informationen eines Diagramms erhalten bleiben.
Eine Alternative zu GraphML stellt das Gexf-Format dar. Es bietet ähnliche Möglichkeiten der Speicherung von Metainformationen wie GraphML, hat jedoch in seiner Definition direkt ein Vokabular für die Darstellung von Netzen integriert.
Ferner kann es Zeitgraphen nativ abbilden, Knoten und Kanten können mit Start- und Endzeitpunkten versehen werden. Somit kann die Zeit definiert werden, in denen Knoten und Kanten existieren. Auch die Angabe mehrerer Zeiträume ist möglich. Vererbungsstrukturen und Hierarchien von Knoten werden unterstützt.[104]
Ein genereller Vorteil von XML ist, das auch Metadaten in XML abgelegt werden können, ohne dass es Probleme beim Parsen der Dateien gibt oder diese durch Syntaxprobleme ungültig werden. Die explizite Auszeichnung und die Möglichkeit der nachhaltigen Dokumentation machen XML-Daten tauglicher als CSV-Daten für Weiterverarbeitung und Langzeitarchivierung. Diese Vorteile werden durch einen höheren Speicherverbrauch in Kauf genommen. Das kann gerade im Web-Umfeld unerwünscht sein, da so höhere Datenraten erzeugt werden.

JSON

Eine ähnliche Ausdrucksstärke wie XML besitzt der JSON-Standard[57]. Er ist weit weniger geschwätzig als XML und hält dafür weniger Möglichkeiten bereit, Metadaten zur Datenstruktur anzugeben. XML ist dabei eher ein Austausch- und Langzeitarchivierungsformat, während JSON auf den konkreten Austausch von Programmen zugeschnitten ist.

RDF und OWL

Ein RDF-Datum, ein Triple, besteht aus Subjekt, Prädikat und Objekt. Aus ihnen lassen sich komplexe Datennetze konstruieren.
Generell kann bei RDF und der Triple Darstellung gesagt werden, dass es sich um eine Menge von gerichteten, gelabelten Kanten handelt, wobei die Labels auf den Kanten je nach Auszeichnung in eine hierarchische Beziehung gebracht werden können.
Hierbei werden Dinge durch URIs identifiziert. URIs können jede Position in einem Triple einnehmen. Dabei sind Subjekt und Prädikat auf jeden Fall durch einen URI zu besetzen. Das Objekt kann auch durch Zahlenwerte oder Zeichenketten besetzt werden. Werte, wie Zeichenketten und Zahlen, bilden immer Blätter, da sie selbst nur in der Objektposition

[57] JavaScript Object Notation[63]

stehen können.
Im RDF-Datenmodell können Attributaussagen und Aussagen über die Graphenstruktur grob getrennt werden. Bei einer Attributaussage würde als Objekt in einem RDF-Triple ein Wert oder eine Beschreibung stehen, aber kein URI. Ein Gegenstand, der durch einen URI bezeichnet wird, kann so beschrieben werden. Im Graphenmodell stellen Subjekte und Objekte Knoten dar und Prädikate die Kanten. Ein "Triple" entspricht dabei einer Kante. Durch seinen Aufbau bildet eine Menge von RDF-Triplen einen Graphen. Dies passiert, wenn URIs in der Objekt- und der Subjektposition mehrfach auftauchen. Dabei kann es durch die Beschreibung der Datenstruktu in den Daten selbst auch zu einem Klassifikationsgraphen werden. Die "rdfs:type" Kanten stellen die Verbindung zwischen der abstakten und der konkreten Datenrepräsentation her.
Das Graphenmodell wird zur Informationsstruktur. Es scheint dabei oberflächlich zu leicht verarbeiten, doch intern wird es komplex, da eine große Menge von einzelnen Aussagen gemacht werden muss, um einen mit anderen Mitteln einfacher abzubildenden Sachverhalt zu beschreiben.
Der Overhead, der hier gemacht wird, ist großer als bei den anderen vorgestellten Formaten und im Gegensatz zu den Strategien zu ihnen vermischen sich hier strukturelle Abbildung und inhaltliche Abbildung.
Es gibt mehrere Abbildungsformate für RDF. Das einfachste Format ist das N-Triple-Format. Nacheinander werden Subjekt Prädikat und Objekt mit Punkt terminiert aufgelistet. Ein etwas komplexeres Format, welches das N-Triple in vollem Umfang mit abbildet, ist Notation3[27].
Notation3 beinhaltet Kurzschreibweisen, die den Overhead, der durch Wiederholungen entsteht, minimieren.
Die N-Triple-Darstellung hat viele Parallelen zur CSV-Darstellung. Es gibt drei Spalten, in denen die Kanten abgelegt werden; zwischen Quelle- und Zielspalte gibt es eine Relationsspalte.
Das RDF-Format kann auch in XML abgebildet werden. Das hat den Nachteil, dass das geschwätzige RDF-Format weiter mit der Geschwätzigkeit[58] von XML aufgeblasen wird. Es macht die Daten auch schwerer lesbar.
In RDF werden beliebige Werte direkt als Teil des Graphen dargestellt. Die Beschreibung der Knoten in RDF mit Attributen wird zwangsweise durch weitere Triple vorgenommen. Kanten können durch die URIs auf den Labels gut nominal gewertet werden und auch in einer Beziehung zueinanderstehen. Kanten können dabei jedoch nicht durch ein Gewicht oder einen numerischen Wert gewertet werden. Dies kann umgangen werden, indem die Kanten durch einen Knoten ersetzt werden und diesem dann Werte zugewiesen werden. Das hat Folgen für die durchschnittliche Entfernung der zu betrachtenden Knoten, da sich die Entfernung dadadurch verdoppelt. Eine RDF-Spezifikation, die das ermöglicht, ist Open Annotation Collaboration.[188]

[58]Durch die menschenlesbaren Steuerzeichen von XML ist die Redundanz von XML sehr hoch.

Durch OWL[59] wird in RDF ein Datenmodell für RDF definiert. Die Beschränkungen, die bei einer relationalen Datenbank durch die Struktur gegeben sind, können durch OWL direkt dokumentiert werden. Die daraus entstehenden Datenmodelle sind allgemein zugänglich, da sie, anders als die Struktur einer relationalen Datenbank, offen liegen.
OWL erlaubt beispielsweise eine komplexe Beschreibung der Kanten- und Knotenklassen. Durch die Einführung einer Hierarchie bzw. Vererbung von Knotentypen gehen die hieraus entstehenden Netze über das Paradigma der Multi-Mode-Netze hinaus, da die Klassen, denen die Knoten angehören, nicht zwangsweise disjunkt sind und daher ein Knoten mehrere Klassen haben kann. Des Weiteren können in OWL alle Klassen auf eine Basisklasse zurückgeführt werden. Dies ist im allgemeinsten Fall das OWL-Thing-Prädikat.
Aber nicht nur die Knoten oder Subjekte können dabei spezifiziert und generalisiert werden, auch Kanten bzw. Kantentypen können in eine Unter- und Oberklassenrelation gesetzt werden.[165]

2.7 Software

Im Folgenden kann nur ein Ausschnitt aus der Vielzahl der Programme und Programmbibliotheken vorgestellt werden, mit denen der Autor im Rahmen dieser Arbeit in Kontakt kam.

2.7.1 Programmbibliotheken

Für die Manipulation von Graphen gibt es eine Vielzahl von Programmbibliotheken, deren Funktionsumfang variiert. Meistens sind die Bibliotheken nur in einer Programmiersprache verfügbar. Die Wahl der Bibliothek ist abhängig von der gewählten Programmiersprache und dem Funktionsumfang.

2.7.2 Programme

Es gibt zur Graphenmanipulation und Visualisierung eine Vielzahl von Programmen mit grafischer Benutzeroberfläche. Hier variieren der Funktionsumfang und die Betriebssystemkompatibilität. Pajek[204] ist ein älteres Programm für Windows, kann aber auch in Verbindung mit Programmen wie Wine auf anderen Plattformen ausgeführt werden. NodeXL[162] ist ein Programm, das direkt an Excel ansetzt. Dies ist insofern interessant, da hier alle Funktionen von Excel zur Datenmanipulation bereitstehen. Dafür ist die kostenpflichtige Basissoftware Voraussetzung.
Im Laufe der Arbeit wurden serverseitige Applikationen erstellt, hierfür wurde Linux verwendet. Desweiteren muss die Software, für eine Automatisierung über eine API angesteuert werden. Daher sind die beiden Softwarepakete nicht verwendet worden.
Es gibt eine Vielzahl von Java-basierten Programmen, deren Vorteil sich aus dem Betriebssystem übergreifenden Einsatzmöglichkeiten ergibt. Hierzu gehören NWB[245], Sci^2[246],

[59] Web Ontology Language, die Abkürzung ist verwirrend, anscheinend wurde mehr Wert auf eine bildliche Abkürzung als auf Logik gelegt.

Gephi[18] und Cytoscape[226].
NWB und Sci^2 sind verwandte Projekte, bei ihnen liegt das Hauptaugenmerk auf dem Errechnen von Metriken und Filtern, die Ergebnisse werden in einem Baum abgelegt. Der Nutzer kann verschiedene Wege der Graphenmanipulationen ausprobieren ohne vorherige Zustände zu verlieren.
Cytoscape kommt aus dem Bereich der Bioinformatik. Cytoscape unterstützt nativ Mehrfachkanten zwischen den Knoten und es gibt eine rege Community und dadurch eine Menge Plug-ins.
Im Rahmen dieser Arbeit wird größtenteils mit Gephi gearbeitet. Gephi kann als Visualisierungsinterface in NWB eingebunden werden. Die Funktionalität von Gephi kann auch über eine API als Bibliothek angesteuert werden. Das ermöglicht, Vorgänge in der Benutzeroberfläche systematisch in einem automatisierten Prozess abzubilden. Gephi unterstützt eine Menge Format Im- und Exporte. Durch die rege Community existieren viele Plug-ins für Gephi, die den Funktionsumfang erheblich erweitern. Gephi hat drei Hauptansichten. In der Visualisierungsansicht, in der hauptsächlich gearbeitet wird, werden Visualisierungsalgorithmen, Färbungen und visuelle Gewichtungen vorgenommen. Auch werden hier Metriken erstellt und Filter definiert.
In der Tabellenansicht für Knoten und Kanten können Gewichtungen und Kategorie- und Texttransformationen vorgenommen werden. Die Dartstellung und Handhabung ähnelt dabei einer Tabellenkalkulationssoftware. Die dritte Ansicht ist die Renderingansicht. In dieser steuert der Nutzer den Export in Vector- und Rastergrafiken.

Kapitel 3

Daten und Netze

In diesem Kapitel wird dargelegt, wie klassische Datenbanken und lokale Datenstrukturen in netztheoretischer Betrachtung visualisiert und explorativ analysiert werden können. Dabei bezieht sich das Kapitel auf Netzstrukturen, die sich in Datenstrukturen und Datenbanken finden lassen. Ein Schwerpunkt liegt dabei auf Relationen aus Strukturdaten und dem wiederholten Auftreten von Entitäten wie in den Abschnitten 3.2.1 und 3.2.2. Es werden aber auch Methoden angerissen, mit denen sich Relationen aus numerischen Eigenschaften herleiten lassen, wie in Abschnitt 3.3 beschrieben wird. Im Abschnitt 3.5 wird die Arachne Datenbank mit automatischen Abfragen untersucht. Dabei wird erläutert, wie sich bipartite Netze durch Projektion in Hinblick auf interessante Muster untersuchen lassen. Dabei werden Kennzahlen und Visualisierungen genutzt um interessante Strukturen zu finden.

3.1 Datenstrukturen und Netze

In diesem Abschnitt wird beschrieben, wie sich klassische Datenstrukturen zu Netzen und netzartigen Beobachtungen verhalten. Netze haben im Verhältnis zu Daten meist ein Problem:
Die meisten Datenstrukturen haben einen linearen Aufbau. So ist jegliche Datenstruktur erst einmal linear.

3.1.1 Listen und Tabellen

Listen oder auch Tupel von Daten sind eines der am weitesten verbreiteten Datenmodelle. Allgemein kann ein Graph einfach als Liste, beispielsweise wie bei einer Kantenliste, abgebildet werden. Das Prinzip der Tabelle oder Matrix ist in Abschnitt 2.6.1 genauer beschrieben. Beim Umwandeln von zueinander zugeordneten Daten zu Graphen können zwei Hauptstrategien verfolgt werden.
Die erste Strategie bezieht sich auf wirkliche Relationen. In einer Tabelle werden zwei Typen von Entitäten beschrieben, beispielsweise Objekte und deren Klassifikation. Daraus lässt sich ein bipartites Netz erstellen.
Die zweite bezieht sich auf Entitäten die einem Wert auffweisen, kann ein Ähnlichkeits-

netz gebildet werden. Das heißt, die Kanten zwischen den Knoten in der ersten Spalte, den Objekten, werden durch ihre Ähnlichkeit im Wert in der zweiten Spalte hergestellt. Dies ergibt ein unipartites Netz. Die Menge der Kanten ergibt sich aus der Bewertung von Ähnlichkeiten. Wenn auch sehr unterschiedliche Werte ähnlich sind, dann entstehen eher viele Kanten, wenn nur sehr nah beieinanderliegende Werte als ähnlich erachtet werden, ergeben sich weniger Kanten.

3.1.2 Geografische Daten

Bei Geodaten handelt es sich um eine spezielle Klasse von Daten, die im geografischen Raum angeordnet werden können. Diese teilen sich allgemein in Vektor- und Rasterdaten. Vektordaten sind Punkte, Pfade und Flächen, die auf einem Ellipsoiden, der Erdkugel, abgebildet werden. Rasterdaten sind Beschreibungen von kleinen, in Raster eingeteilte Flächen, die durch Werte beschrieben werden. Dies entspricht einer normalen Rastergrafik, wie sie für Bilder verwendet werden mit dem Unterschied, dass diese Raster auf der Oberfläche eines Ellipsoiden abgebildet werden.
Im Allgemeinen kann die Verarbeitung der meisten Geodatensätze auch auf Daten mit Anordnungen im zweidimensionalen Raum angewendet werden. Dabei ist die Einschränkung, dass es sich um reale Daten in einem Raum mit klar definierten Distanzen handelt, sehr wichtig. Aus den genannten Vektordaten lassen sich Netze bilden. Dabei ist die Zuordnung von Formen zu identifizierbaren Einheiten die Grundidee einer Graphenverarbeitung.
Punkte sind die simpelste Form von Geodaten, sie können durch eine einzige Koordinate im Raum beschrieben werden. Punkte können je nach Modell automatisch oder nach spezifischen Gesichtspunkten verbunden werden. Hier können beispielsweise die am nächsten beieinanderliegenden Punkte als Knoten verbunden werden. Dadurch ergibt sich ein Nearest-Neighbour-Graph. Alternativ kann beispielsweise auch eine Delaunay Triangulation verwendet werden, um aus Punkten einen Graphen zu bauen.
Pfade sind Abfolgen von Punkten. Aufeinanderfolgende Knoten werden durch Kanten verbunden. Kreuzungen zwischen Pfaden können Kreuzungsknoten bilden. So lassen sich aus mehreren überkreuzenden Pfaden Graphen errechnen. Die Pfade zwischen den neuen Knoten, die meist anhand von verbundenen oder geordneten Punktemengen entstehen, können dabei nach Belieben, solange sie keine Schnittmenge darstellen, zu gewichteten Kanten zusammengefasst werden. Die Zwischenknoten müssen nicht zwangsläufig ins Graphenmodell übernommen werden.
Bei den Flächen können gemeinsame Grenzen oder Überlappungen als Verbindungen gesehen werden. Flächen sind dabei die Knoten, die Überlappungen und Angrenzungen bilden die Kanten. Dieses Prinzip wird zum Beispiel bei der gestaltung von Spielfeldern wie im Spiel Risiko[1] verwendet.[85]
Gazetteere sind Ortslexika, sie stellen strukturierte Informationen zu bezeichneten Orten bereit. In Gazetteren können je nach zweck verschiedene Ortssysteme abgelildet sein,

[1] Siehe Abschnitt 2.2.4.

beispielsweise Verwaltungstechnische Strukturen. Sie sind häufig hierarchisch strukturiert, daher bildet sich eine baumartige Struktur, deren Wurzel das Universum oder die Erde der alles verbindende Punkt ist.
Informationen wie sie in klassischen Gazetteeren vorkommen können als teil von anderen Standards, wie beispielsweise im CIDOC-CRM[76] abgebildet werden.[2]

Visualisierung

Anders als bei Graphen-Visualisierungen aus abstrakten Ideen wie Kommunikation, Freundschaft oder Ähnlichkeit ist die Struktur, die aus geografischen Netzwerken hervorgeht, immer an die Position gebunden. Es besteht also immer die Möglichkeit, geografische Positionen in die Visualisierung einzubeziehen. Daraus bildet sich eine nachvollziehbare und überprüfbare "richtige" Grundordnung, die auf konzeptioneller Seite reproduziert werden lassen kann. Andererseits kann sie auf der konzeptionellen Seite auch eine andere Ordnung annehmen als die geografische Ordnung.

3.1.3 Text

Eine der am wenigsten strukturierten Datenformate ist der Freitext. Dabei kann es sich, je nach Betrachtungsweise, um ein sehr simples sowie eines der komplexesten Datenformate handeln.
Einerseits ist es eines der simpelsten Datenformate, da keine Regeln und keine formalen Strukturen eingehalten werden müssen.
Andererseits ist es das komplexeste Datenformat, da für die Interpretation Kontext vorrausgesetzt wird und so gut wie alles mit Text beschrieben werden kann.
Diese Paradoxie kann damit entschärft werden, dass man Text als die komplexeste Datenstruktur sehen kann, die vom Menschen "verarbeitet" werden kann, jedoch nur sehr begrenzt im Computer. Viele Publikationen sind nach gewissen Standards geschrieben und folgen einer einheitlichen Rechtsschreibung, Formatstandards, Zitierrichtlinien und verwenden Inhaltsverzeichnisse. Andere Texte wie Kommentare in sozialen Medien werden nicht in dieser Hinsicht redigiert und dadurch strukturell normalisiert. Da es sich bei der Analyse und Visualisierung aber immer um eine Simplifizierung handelt, wird versucht, wichtige Zusammenhänge mehr oder weniger automatisch aus dem Freitext zu extrahieren.

Graphen aus Freitext

Ein Graph wird aus Knoten und Kanten aufgebaut und diese Knoten und Kanten müssen Inhaltlich bestimmt werden. Die Knoten können durch das Identifizieren von klar eingrenzbaren Identitäten benannt werden. Das ist nicht trivial und es benötigt Aufwand, Entitäten in Texte zu identifizieren und voneinander abzugrenzen. Selbst Namen von Personen und Organisationen lassen sich nicht ohne Aufwand fehlerfrei identifizieren. Die

[2] Wie CIDOC-CRM verwendet werden kann wird in Abschnitt 4.2.3 genauer besprochen.

einfachste Möglichkeit, identifizierbare Knoten zu schaffen, ist das Tokenizing. In seiner simpelsten Form ist ein Token eine Unterzeichenkette, die zwischen Satzzeichen und/oder Leerzeichen steht. Das berücksichtigt jedoch nicht das Vorkommen von Entitäten, die durch verschiedene Worte oder durch eine Kombination aus mehreren Worten identifiziert werden.
Sind die Knoten definiert und extrahiert müssen Relation bestimmt werden. Die Variation hat immer mit dem gemeinsamen Auftreten von Knoten zu tun. Eine der einfachsten Methoden ist es, Wortfolgen, Sätze oder Textabschnitte zu analysieren, in denen Wörter zusammen vorkommen und aus diesen Kanten bilden. An dieser Stelle würde sich ein bipartites Netz bilden. Die Quellknoten wären Absätze, Abschnitte oder Sätze und die Zielknoten wären Tokens, Wörter oder erkannte Entitäten.
Die automatische Extraktion von strukturierten Daten aus Text fällt unter den Oberbegriff Natural Language Processing kurz NLP. Für die Extraktion von Graphen ist das Erkennen von Entitäten und Relationen nötig. Viele Methoden aus der künstlichen Intelligenz benötigen eine vielzahl an Trainingsdaten, um in einem Textkorpus Strukturen zu erkennen. Dazu kommt, dass die Methoden nur begrenzt zuverlässig sind, was die daraus entstehenden Daten anfällig für Fehler macht.
Eines der größten Probleme in diesem Zusammenhang sind False-Positives, also Entitäten, die fälschlicherweise erkannt wurden. Die daraus entstehenden Fehler können Netzwerk-Maßzahlen stark verfälschen.
Die Identifikation von Entitäten in Texten kann anhand von Spezialwörterbüchern, Listen von Akteuren und ähnlichen Zielmengen erfolgen. Hier hilft die Semistrukturiertheit von Begriffen und Entitäten, über die geredet wird. In Veröffentlichungen von Artikeln zur Genetik können beispielsweise Krankheiten und Gene anhand ihrer Bezeichnungen recht einfach identifiziert und in Beziehung gebracht werden.[54]
Diese automatischen Verfahren werden als Named Entity Recognition(NER) oder Named Entity Classification(NEC) bezeichnet. Es werden dabei Dinge (NER) oder Konzepte (NEC) in Texten identifiziert und ausgezeichnet. Diese identifizierten Entitäten können auch als Datensatz beschrieben sein.[179]
NER und NEC-Systeme können per Hand, anhand von Regelwerken[3] oder automatisch erstellt werden. In neuerer Zeit wurden Lernverfahren aus dem Bereich der künstlichen Intelligenz zu diesem Zweck immer beliebter.[178] Diese Lernverfahren benötigen jedoch einen annotierten Korpus von Texten und eine Liste von Entitäten, die erkannt werden sollen.
Bei neueren Verfahren handelt es sich meist um überwachte Lernverfahren, die auf verschiedenen Ansätzen der künstlichen Intelligenz beruhen können. Unüberwachte Lernverfahren werden auch verwendet, schneiden in Tests jedoch meistens noch schlechter ab.[144] Automatische Verfahren müssen nicht zwangsweise für sich alleine stehen. Sie können auch den Menschen beim Auszeichnen von Texten unterstützen.[239] Vorgehen dieser Art gibt es in der Soziologie, dort werden Texte in Hinblick auf die Nennung spezieller Akteure und

[3] ein solches Vorgehen wird in Abschnitt 3.2.2 vorgestellt.

deren Interaktionen analysiert.[74]
Die Extraktion von Relationen hängt von der Zielmenge der Relationen ab. Am simpelsten aber auch aussagelosesten sind gemeinsames Vorkommen von Wörtern direkt nacheinander, in einem Satz, einem Abschnitt, einem Kapitel, einem Buch, einem Beitrag in einem sozialen Netzwerk oder einer Internetseite.
Abstrakte Zusammenhänge lassen sich direkt aus der Struktur herleiten. Explizite semantische Beziehungen müssen jedoch mit mehr Aufwand extrahiert werden. Relationen wie Feindschaft zwischen Personen, auseinander hervorgehende Ideen und künstlerische Anlehnung zwischen Werken brauchen ein Vokabular an identifizierbaren Relationen. Desweiteren werden Methoden benötigt die definierten Verbindungen aus Freitext zu extrahieren, beispielsweise mit hilfe der die Satzstruktur.[206]

3.1.4 XML als strukturierter Text

XML ist ein etablierter Auszeichnungsstandard für Text und Strurierte Daten. XML eignet sich zum hierarchischen Ablegen von Informationen sowie zum Annotieren von Texten. Daher haben sich viele geisteswissenschaftliche Projekte gefunden, die XML verwenden, um die Ergebnisse ihrer Arbeit zu codieren. Einen Standard zum Auszeichenen von Text in XML bietet die TEI (Text encoding initiative)[247].
Die Auszeichnung von Text ist eine Möglichkeit, Freitext vorzuverarbeiten, beispielsweise Entitäten zu markieren. Text in seiner unstrukturierten Form ist in seiner genauen Verarbeitung komplexer, da er offener gestaltet werden kann.
Da auch viele Texte in HTML vorliegen, können sie mit den gleichen Mitteln verarbeitet werden wie XML. Dabei ist der Schritt von HTML zu XHTML zu gehen, um problemlos XML-Werkzeuge verwenden zu können.

Struktureigenschaften und Einschränkungen

XML besteht aus ineinander verschachtelte Einheiten sogenannten Tags. XML bildet strukturell einen Baum.
Daten, die im Rahmen dieses Kapitels untersucht werden, beruhen auf Text, der mit XML ausgezeichnet wurde. Der annotierte Text kann fortlaufend als Text gelesen werden. Die andere Anwendungsmöglichkeit für XML ist es strukturierte Daten abzulegen. Beim Abbilden von Daten gibt es keinen fortlaufenden Text, sondern klar abgegrenzte Einzelwerte, deren Reihenfolge meistens keine Rolle spielt.
XML kann auf Text immer nur eine einzige Baumstruktur auf einmal abbilden. Tags können jedoch nicht übereinanderliegen, sodass sich diese Verschachtelungen überschneiden. Das würde einen Bruch in der Baumstruktur bedeuten. Alternative Strukturen können per Verweis mit Identifiern realisiert werden.
Als Beispiel kann die klassische Einteilung von Texten in Seiten, Kapitel, Abschnitte und Absätze dienen.
Diese Strukturen lassen sich gut in XML abbilden, da sich die Einteilungen in Kapitel und Absätze nicht überschneiden. Problematisch wäre eine Abbildung von sich überschneidenden Ordnungen. Wenn die Seiten und Kapitel in einer Printpublikation in einem Tag einge-

schlossen sind und dann auf ein und derselben Seite ein Kapitel endet und ein neues beginnt. Das kann nicht mit einem Tag abgebildet werden, da sich zwei Tags überlappen und damit die klare Struktur brechen würden.
Eine Lösung für dieses Problem wird im TEI-Standard behandelt. In TEI werden Strukturelemente mit dem **div** Tag ausgezeichnet das **type** Attribut typisiert den Abschnitt beispielsweise in **chapter** oder **section**. Um einen Abschnitt zu unterbrechen und fort zu setzen wird mit einem **n** Attribut die Nummer des Abschnitts angegeben.
Auch Graphen lassen sich als XML abbilden, jedoch verliert das XML dabei seinen linearen Charakter, wie er beim Fließtext gegeben ist. Auch vollständige Graphen mit Kreisen sind keine Bäume und lassen sich daher nicht in der nativen Struktur von XML abbilden.Eine Lösung, wie Netzdaten in XML abgebildet werden können, ist GraphML[4]. Es gibt in GraphML wie auch in anderen Formaten eine Aufteilung in zwei disjunkte Mengen, den Knoten und den Kanten. Die Kanten beziehen sich dabei immer auf die definierten Knoten. Eine Verbindung dieser beiden Teile über die Datenstruktur ist nicht möglich. Daher werden, wie bei TEI, Identifier für die Verweise zwischen Knoten und Kanten verwendet. Die Attribute der einzelnen Knoten und Kanten können die Vorteile einer XML-Struktur nutzen. Auf diese Weise lassen sich recht flexibel einzelne Werte hinzufügen.

Von XML zum Graphen

XML ermöglicht es, Dinge in Texten zu markieren. Somit ist XML ein möglicher Repräsentationsschritt für die Ergebnisse von NER- und NEC-Verfahren. Ein Wort, das einen bestimmten Gegenstand beschreibt, kann in einem Tag eingeschlossen werden. Dadurch ist es in Verbindung mit einem Tag direkt semantisch ausgezeichnet und ist so beispielsweise von Synonymen abgrenzbar. Mithilfe der Attribute kann ein Verweis auf einen Identifikator wie etwa einen URI, eine ISBN, einen Paragrafen oder eine Personenidentifikationsnummer hinterlegt werden.
Auch können verschachtelte Tags direkt den Rahmen für ein bipartites Netz bilden:
Die erste Art von Tag bildet dabei die Knoten der ersten Kategorie, im Weiteren auch das umschließende oder äußere Tag genannt.
Die Knoten der zweiten Kategorie befinden sich in den umschließenden Tags. Sie werden innere Tags genannt.
Die Struktur von XML bietet den Vorteil, dass diese Zuordnungen genau definiert werden können. Würden sich die Tags überschneiden, wäre eine solch genaue Zuordnung nicht möglich. Die Überschneidungsfreiheit löst hier eine Problematik der Granularität, da kontrolliert werden kann, ab welcher Ebene eines XML-Baums etwas verarbeitet wird.
Modellieren lässt sich dieser Vorgang anhand von zwei XPath-Befehlen, die in zwei Schleifen durchlaufen werden. XPath ist dabei eine Beschreibungssprache, um einen oder mehrere Tags oder Eigenschaften von Tags in XML Dateien systematisch zu beschreiben.
Im Folgenden wird ein Algorithmus skizziert, der auf Grundlage der Beschreibung des

[4]Weitere Informationen zu GraphML finden sich in Abschnitt 2.6.2.

äußeren und inneren XPath-Ausdrucks eine Kantenliste erstellt. Dabei kommen alle Knoten der ersten Art auf die Quellknoten-Position und alle Knoten der zweiten Art auf die Zielknoten-Position der Kanten.
Sollte der äußere XPathausdruck nicht vom Root-Element ausgehend definiert werden, dann muss sichergestellt werden, dass er nicht auf verschachtelte Elemente matcht. Ansonsten können Granularitätsproblemen entstehen.

Listing 3.1: *Algorithmus zur Graphenextraktion*

```
Xpath : XPOut
Xpath : XPIn
XMLDocument : XDoc
List of Edges : EdgeList

for XDoc.getAll(XPOut) as Out
        for Out.getAll(XPIn) as In
                EdgeNew = new Edge -> Souce=Out, Target=In
                EdgeList.append(EdgeNew)
```

Beim beschriebenen Vorgang hängt die Bildung eines Netzes davon ab, ob die Elemente, die durch den inneren XPath Ausdruck beschrieben werden, auch in mehreren äußeren Elementen vorkommen.
Dieser Algorithmus lässt sich einfach erweitern, um mehr Modelleigenschaften erfassen zu können.
Um beispielsweise die Verbindung zu typisieren, kann ein dritter XPath-Ausdruck hinzugefügt werden. Dieser kann ausgehend vom inneren Element ein weiteres Element beschreiben, dass die Beziehung spezifiziert.

Listing 3.2: *Algorithmus zur Graphenextraktion mit typisierten Kanten*

```
Xpath : XPOut
Xpath : XPIn
Xpath : XPSem
XMLDocument : XDoc
List of Edges : EdgeList

for XDoc.getAll(XPOut) as Out
        for Out.getAll(XPIn) as In
                Sem = In.getFirst(XPSem)
                EdgeNew = new Edge -> Souce=Out, Target=In, Type=Sem
                EdgeList.append(EdgeNew)
```

Das gleiche Vorgehen ist auch auf HTML-Dokumenten möglich. Hier werden statt XPath Ausdrücken CSS Selektoren genutzt. Der Vorteil von HTML ist, dass der Laie noch einfacher wiederkehrende Positionen und Strukturen visuell erkennen und markieren kann. In heutigen Browsern sind Einwicklerwerkzeuge zum Untersuchen von HTML-Elementen

enthalten. Diese können CSS Selektoren automatisch bestimmen. Ähnliche Funktionen gibt es auch in XML-Editoren.

3.2 Anwendungsbeispiele für XML

Im Folgenden wird anhand einiger Beispiele dargestellt, wie sich aus XML-Dateien Graphen extrahieren lassen. Dabei geht es meist um das Erlangen des Einblicks in die Struktur von Datensätzen.
In den zwei Beispielen sind Datensätze in XML codiert und werden in Netze überführt. Auf diese Weise wird versucht, einen Überblick über die Inhalte zu schaffen und einfache strukturelle Fragen zu beantworten.

3.2.1 Macbeth

Ein Anwendungsgebiet von TEI ist die Annotation von Bühnenstücken. Hierzu ist ein Korpus von Daten zu Shakespeare vorhanden.
Im Folgenden werden die Dokumente aus der Beispielapplikation von eXist verwendet[225]. Diese gehen auf das WordHord Projekt der North Western University mit der Perseus Digital Library zurück.[269]
Die Funktion des Programms konzentriert sich auf Textanalysen der Werke.

Computergestützte Verarbeitung von Shakespeares Werken

Eine Netzwerkvisualisierung zu Shakespeares Stücken gibt es von Paul Mutton[177]. Er hat die Stücke mithilfe eines Tools zur Visualisierung von IRC-Chats[5] visualisiert. Durch die Datenlage lassen sich dynamische Zeitnetzwerke aus Chats visualisieren. Dafür werden die aufeinanderfolgenden Sprechrollen verlinkt.[6]
Andere Untersuchungen zu Stücken Shakespeares legen nahe, dass nur eine begrenzte Menge von Charakteren, etwa 30 bis 40, in diesen Stücken mitspielen kann, damit ein Stück für den Rezipienten übersichtlich bleibt und er dem Stück folgen kann. Auch sind die Entfernungen im Sinne des Small-World Durchmessers nicht sehr hoch. Jeder Charakter ist durchschnittlich nur über einen anderen Charakter mit allen anderen verbunden.[240]

Konverstationsnetzwerke

Andere Arbeiten mit inhaltlichem Anspruch wie die von Elson et al.[86] stellen Kriterien für ein “conversational network” zum Thema “Literary Fiction” auf[86]:

1. The characters are in the same place at the same time;
2. The characters take turns speaking; and
3. The characters are mutually aware of each other and each character’s speech

[5] Das hierfür verwendete Tool heißt Piespy[176].

[6] Eine solche Visualisierung für Macbeth befindet sich unter `http://www.tux.org/pub/sites/jibble.org/shakespeare/divx/macbeth.avi` zuletzt gesehen am 03.11.2015.

is mutually intended for the other to hear.

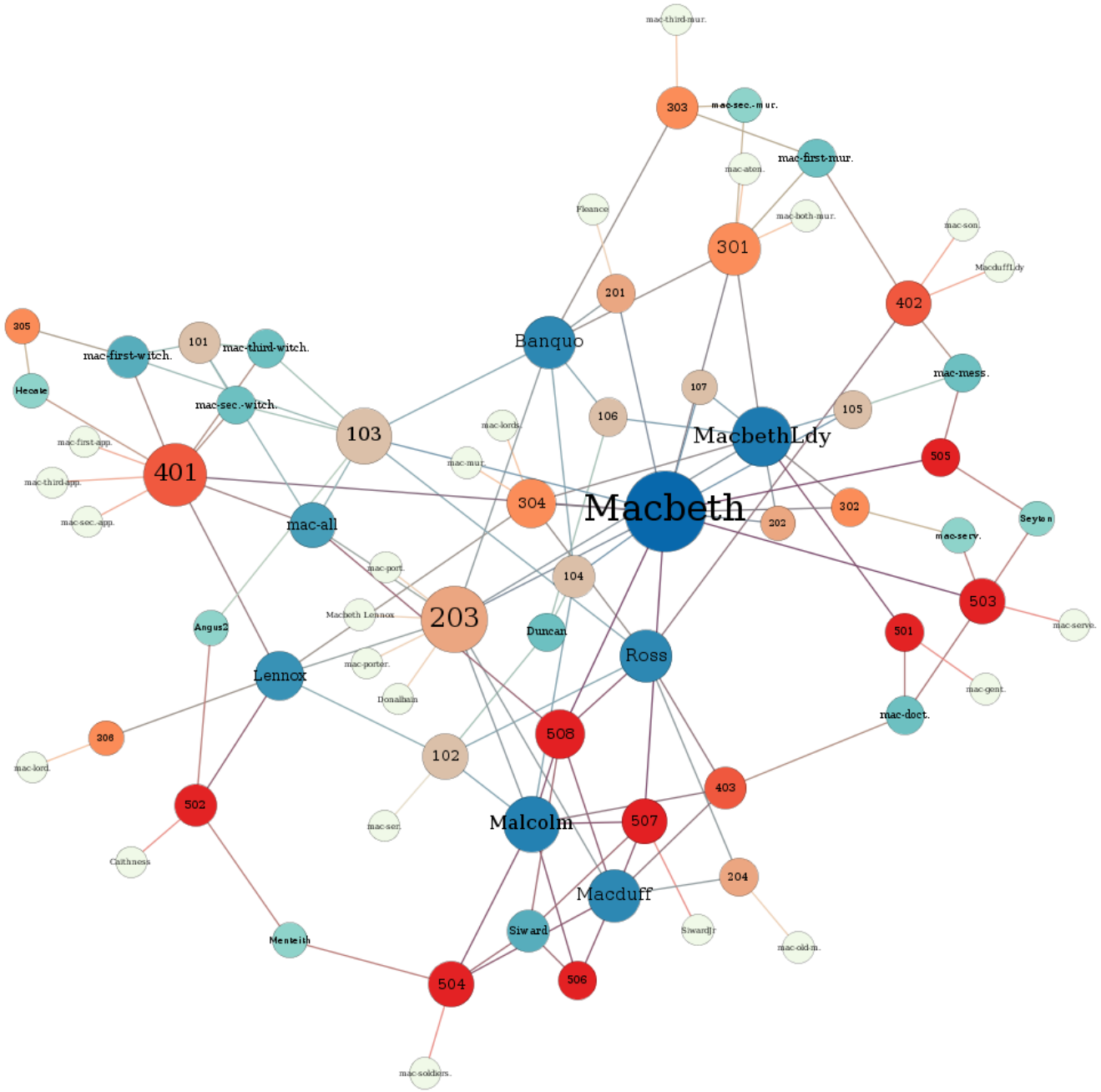
Nach dieser Definition sind Konversationen Austausch. Das lässt sich auf die Datengrundlage zurückführen, in ihrer Untersuchung wurden Romane[7] behandelt. Hier ist eine definitorische Abgrenzung der direkten Rede vom restlichen Text wichtig. Daher stellen die Kriterien auch eine technisch erfüllbare Definition dar.
Das erste Kriterium wurde auf den Bühnen der Shakespearezeit meist erfüllt. Lokalisierung oder zeitliche einordnung wurden hauptsächlich durch den Stücktext vorgenommen. Für Bühnenstücke generell ist die strikte verwendung des Orts schwieriger. Die Bühne als Ort kann potenziell parallel handelnde Szenen aufzeigen. Auch sind so Telefongespräche, IRC Chats sowie Visionen und Träume nie Teil des Netzes, außer man legt den Begriff "place" sehr weit aus. Das dritte Kriterium schließt flüsternde Personen genauso wie im Verborgenen beobachtende Charaktere aus. Es wäre ein Netz des gemeinsam einverständlichen Austauschs. Es ist ein Teil-Informationsnetz, da asynchrone Informationen wie beobachteter Geheimnisaustausch nicht einfließen würde.
Die Gewichtung der Knoten erfolgte bei Elson et al. nach der Menge der Konversation und die der Kanten gemäß der gemeinsamen Kommunikation der durch die Knoten repräsentierten Charaktere. Dabei fällt auf, dass so die zentralen Charaktere immer immens gewichtet sind und die Visualisierungen dominieren.[86]

Inhalt und Strukturen

Die folgenden Analyse bezieht sich auf ein bipartites Netz aus Akteuren in den jeweiligen Szenen, in denen sie sprechen. Daraus können Netzvisualisierungen erstellt werden, wie in Abbildung 3.1 zu sehen ist. Eine Gewichtung der Kanten wurde nicht vorgenommen. Die Kanten werden als ungerichtet betrachtet. Die Charaktere sind mit ihrer ID aus dem TEI-Dokument beschriftet. Die Szenen folgen einem einfachen Schema, das sich auch aus deren IDs erschließt: die erste Ziffer gibt den Akt an, die folgenden Ziffern die Szene. Nummer 203 ist also die dritte Szene im zweiten Akt. Die Szenen haben einen Farbgradienten von Weiß nach Rot, abhängig von ihrer Nummer, weiße Szenen sind am Anfang des Stückes, rote Szenen am Ende. Charaktere sind nach Ihrer Wichtigkeit für das Stück (der Grad) mit einem Frabgradienten von Weiß nach Blau gefärbt.
Als zentralster Knoten mit den meisten Kanten ist hier ganz klar die Hauptfigur Macbeth, zu nennen. Nach ihm ist das Stück benannt. Wichtig ist zu wissen, dass der zentrale Knoten den Charakter und nicht das gleichnamige Stück beschreibt! Das ist in multipartiten Netzen nicht immer selbstverständlich.
Die visualisierte Anordnung der Szenen und Charaktere zeigt eine nicht chronologische Sortierung der Szenen mit den darin vorkommenden Akteuren als Ordnungselement. Eine lineare Anordnung der Szenen nach ihrem Zeitpunkt wird durch die Zuordnung zu Akteuren nicht erzeugt.

[7] Engl.: novels

Abbildung 3.1: *Visualisierung aller Szenen und Charaktere in Macbeth*



Die Hexen nehmen in dieser Visualisierung keine zentrale Position ein. Sie sind Sichtbar abseits positioniert auf der linken Seite der Visualisierung. Zu beginn in der dritten Szene des ersten Akts (103) der prohezeien sie Macbeth das er der neue König werden kann. Die Prophezeiung ist Auslöser der Tragödie und daher inhaltlich zentral.
Weiterhin sind Szene (305) und (101) sichtbar abseits positioniert sie bilden das "backstage"[8] der Hexen im Stück. In der fünften Szene des dritten Akts (305) besprechen sie weitere Prophezeiungen die Macbeth in die Irre leiten sollen und so ins tragische Ende führen. In der ersten Szene des vierten Akts (401) prophezeien sie Macbeth, dass er nur von Jemandem besigt werden kann der von einer Frau geboren wurde. Das wägt ihn in Sicherheit.

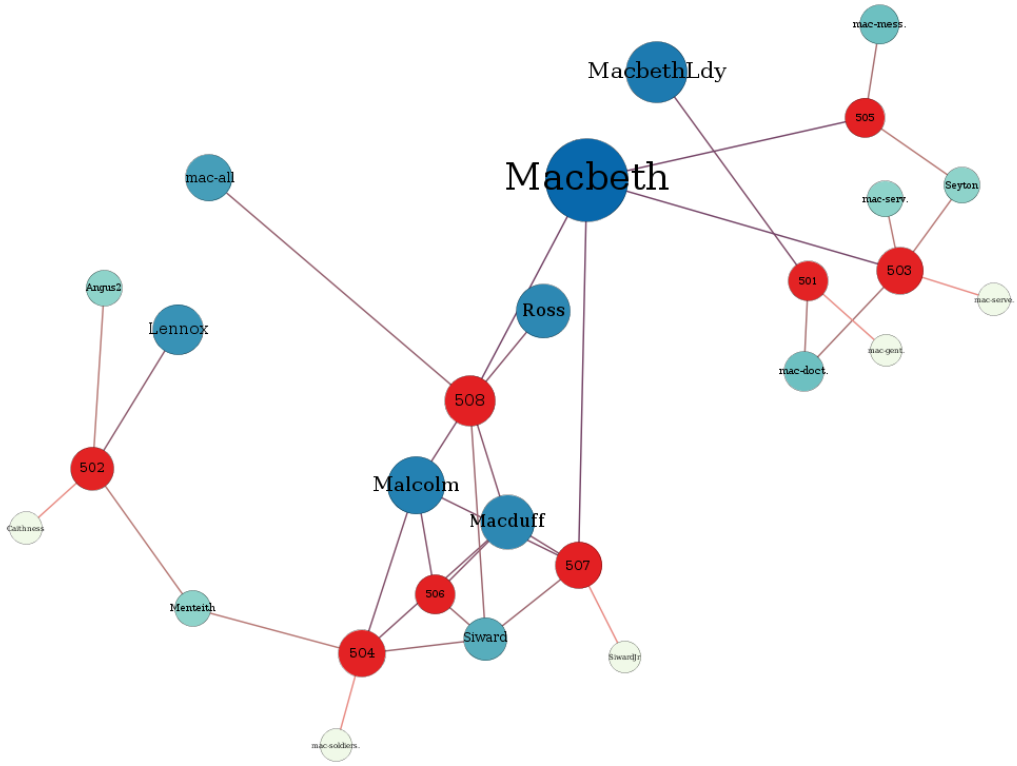
Im Weiteren werden die Auftritte des fünften Akts genauer untersucht, wie sie in Abbildung 3.2 gefiltert gezeigt werden. Die Auftritte des fünften Akts zeigen die Spaltung in zwei Lager. Shakespeare verschränkt die Szenen um Macduff (504,506) und Macbeth (503,505). Macduff zieht mit Malcolm, dem Sohn des von Macbeth ermordeten Königs Ducan, gegen Macbeth. Die letzten Szenen vereinigen die beiden Parteien dann im finalen Kampf miteinander (507,508).
Im nächsten Satz befindet sich ein sogenannter Spoiler, der das Ende und damit das Spannungselement der Geschichte vorwegnimmt. Macduff gewinnt im Kampf gegen Macbeth, er kam per Kaiserschnitt auf die Welt und ist daher auch laut der Prophezeiung in der lage Macbeth zu besigen.
Ein klar sichtbares Detail ist, dass Lady Macbeth nur mit der ersten Szene (501) verknüpft ist. In dieser nimmt sich Lady Macbeth das Leben, da sie Macbeth zum Mord an König Ducan verleitet hat und von ihren Schuldgeführen in den Wahnsinn getrieben wurde. Diese Information wird Macbeth von einem Artzt in der dritten Szene des fünften Aktes mitgeteilt(503). Dieser Arzt stellt im fünften Akt strukturell die einzige Verbindung zwischen Macbeth und Lady Macbeth dar.
Eine für die Handlung bedeutende Szene ist die dritte des zweiten Aktes (203). Sie hat in der Visualisierung eine Zentrale position. Das Auffinden des von Macbeth getöteten Königs Ducan geschieht unter den Augen aller zentralen Personen im Stück. Diese Szene daher ist im informationstechnischen Sinne eine Schlüsselszene. Sie ist von Grad und Betweenness Centrality die am stärksten gewichtete, daher auch technisch, die zentralste Szene.

Ausfallende Schauspieler

In größeren Produktionen werden in der Probephase mehrere Szenen parallel geprobt. Das setzt voraus, dass sich der Einsatz der Schauspieler in einzelnen Szenen nicht überschneidet. Dieses Kriterium kann als Filter genutzt werden. Ist ein Schauspieler krank oder verreist und hat keinen Ersatz, kann anhand einer Graphendarstellung visuell Vermittelt werden,

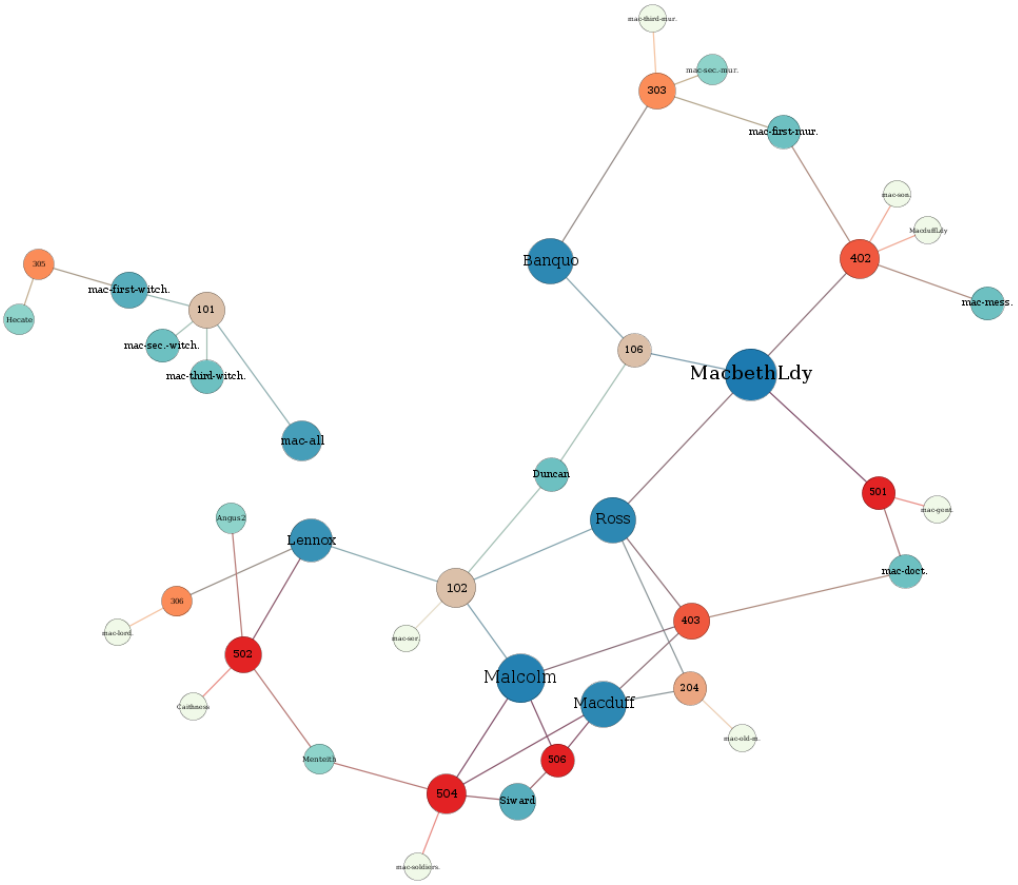
[8]Hier angelehnt an den Begriff nach Goffman.[107] Es ist der soziologische Rückzugsraum gemeint in den nur Hexen Zutritt haben und nicht das Backstage eines wirklichen Schauspielhauses. Hier können die Hexen offen ihre Verschwörung gegen Macbeth besprechen.

Abbildung 3.2: *Filter auf 3.1 mit Fokus auf den fünften, finalen Akt*



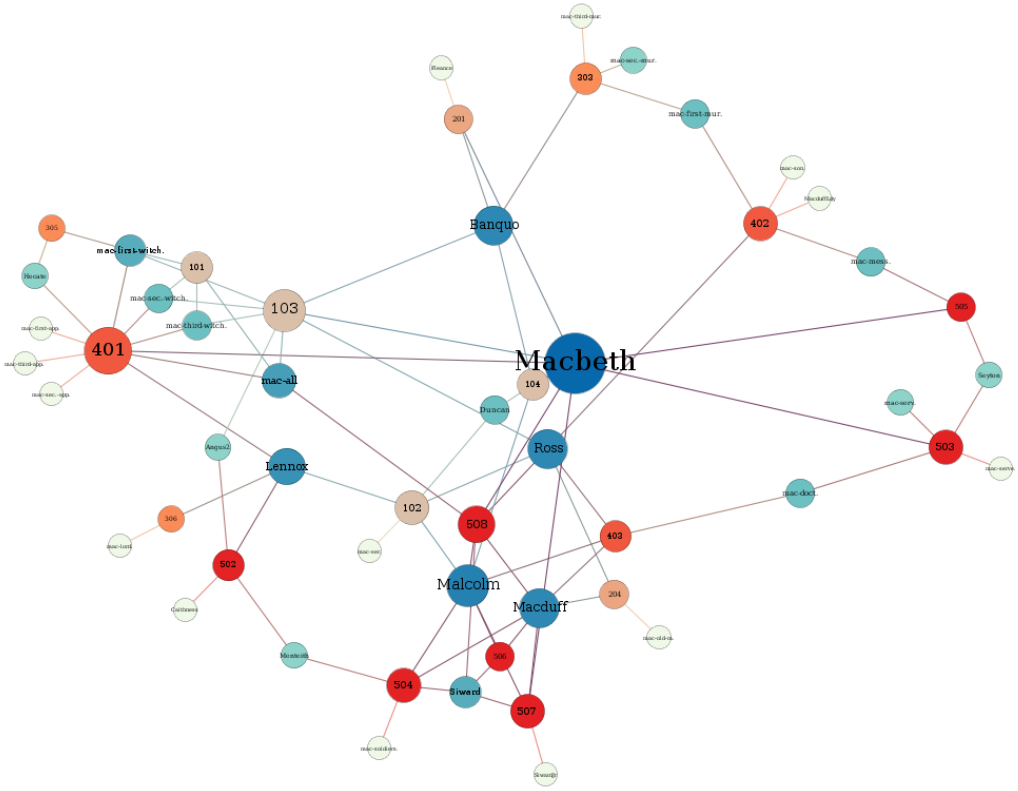
welche Szenen für Proben übrig bleiben. Hier im Vordergrund steht ein Ausfallen der zentralsten Figur, der des Macbeth. Wie Abbildung 3.3 zeigt, fallen mit dem Ausfallen von Macbeth viele Szenen heraus, die geprobt werden können. Ohne die zentrale Figur zerfällt das Stück des Weiteren in mehrere Komponenten.
Weniger kritisch ist die Probelage für die Regie, wenn Lady Macbeth ausfällt. Dann können auch weiterhin viele Szenen geprobt werden, siehe Abbildung 3.4. Das gilt vorallem für den fünften Akt wie auch für die Szenen mit den Hexen.
Der hier angewendete Filter entfernt alle Knoten, mit denen ein Schauspieler verbunden

Abbildung 3.3: *Abbildung wie 3.1. Macbeth und alle Szene in denen er mitspielt sind herausgefiltert.*



ist, der gefiltert wird. Das sind durch die Bipartät des Graphen nur Szenen, da es keine Kanten zwischen den Schauspielern gibt.
Sollten dabei andere Schauspieler frei werden und keine Verbindung mehr haben, werden auch sie herausgefiltert.
Die Aufstellung dieser herausgefilterten Figuren mit einem Grad von null können beim Ausfall des Darstellers von Macbeth nicht mehr proben. In diesem Fall sind 15 Rollen be-

Abbildung 3.4: *Abbildung wie 3.1. Lady Macbeth und alle Szene in denen sie mitspielt sind herausgefiltert.*



troffen, jedoch handelt es sich meistens um Nebenrollen und Statisten. Bei Lady Macbeth sind es 8, auch hier sind größtenteils Nebenrollen betroffen.[9]

Visualisierung

In den Darstellungen 3.1, 3.2, 3.3 und 3.4 wurde nur gefiltert, die Positionen wurden ansonsten nicht verändert, um die Orientierung des Betrachters zu erhalten. Zur initialen Positionierung wurde ein einfacher, kräftebasierter Algorithmus verwendet.

[9]Der Charakter "mac-all" wurde nicht mitgezählt da er kein Charakter, sondern eine variierende Anzahl von Charakteren und Statisten abbildet.

Technische Realisierung

Die Akteure konnten mithilfe von XPath-Ausdrücken aus dem Dokument extrahiert werden. Der Ausgangspunkt im XML-Dokument waren die einzelnen Auftritte mit Text. Hier wurden die Sprecher durch einen XPath-Ausdruck extrahiert. Durch das Referenzieren von vorherigen Elementen im Baum wie der Szene oder des Aktes können die Strukturdaten in Netze überführt werden. Zu den referenzierten Abschnitten wurden auch reguläre Ausdrücke verwendet, um sinnvolle und lesbare Teilstrings zu erzeugen.
Die Überführung in Netze setzt zu dem Zeitpunkt ein, in dem all diese Daten zusammengetragen werden, dann werden die Verbindungen definiert. Hier wurde bestimmt, dass alle gefundenen Knoten mit einer Typisierung ausgezeichnet werden. Diese Knoten werden dann miteinander verbunden. Dabei werden die Kanten durch Kombination der Typen der Ausgangsknoten typisiert.

Generalisierbarkeit

Wenn diese Auszeichnungsuntermenge von TEI einmal erfasst wurde, dann können auch andere Stücke in der gleichen Auszeichnung vom gleichen Programm in Graphen umgewandelt, analysiert und visualisiert werden.
Durch die einheitliche Strukturierung können auch verschiedene Versionen und Abänderungen des gleichen literarischen Stoffs verglichen werden. Hier ist eine Frage, inwiefern neue Auszeichnungsstile nicht vielleicht auch eine Anpassung der Software notwendig machen können. Die angeführten regulären Ausdrücke, welche Informationen aus den IDs extrahieren, können bei einer anderen ID-Vergabe-Praxis gegebenenfalls aufhören Ergebnisse zu liefern. Hier könnte alternativ das Tag **castList** ausgewertet werden. Es enthält eine Textbezeichnung aller Charaktere mit ihren IDs.
Eine automatische Visualisierung aller Stücke wäre möglich, dazu müsste eine Kombination aus Parametern für die Algorithmen gefunden werden, die sich an Parametern der Werksstrukturen orientiert. Dabei ist die Parametrisierung dringend notwendig, um lesbare Ergebnisse und keine “Hairballs” zu erzeugen.
Generalisierbarkeit würde sich an einer Angleichung der Parameter des Stücks bzw. des Graphen orientieren. Damit würden vergleichbare Visualisierungen geschaffen. Bei diesem Vorgehen handelt es sich eher um ein Kochrezept als um eine strenge Wissenschaft. Das Rezept würde beispielsweise besagen, wie viel Abstoßung zwischen den Knoten und wie viel Anziehung auf den Kanten liegen sollte. Das Rezept lässt sich dann wie ein Koch- oder Backrezept an der Anzahl der zu Bekochenden oder hier der Personen und Szenen sowie an der Anzahl ihrer Verbindungen hochrechnen. Die Güte dieses Ansatzes ließe sich an den vielen automatisch erstellten Visualisierungen testen.[10]
Unklarheiten bei den Entitäten hinter den Knoten, sollten aufgrund der eindeutigen Iden-

[10] Hier kommt wieder das Problem auf, dass bei Visualisierungen sehr viel Zeit für das Betrachten und Bewerten verwendet werden muss. Hierbei können gängige Methoden der künstlichen Intelligenz nur bedingt helfen. Generell wäre es jedoch möglich, ein neuronales Netz zu trainieren, das die abstrakte Eigenschaft der “Lesbarkeit” bewertet. Sollte so etwas zuverlässig gelingen, könnten die Anpassungen der Parameter des Algorithmus zum Netz auch vom neuronalen Netz, der KI, vorgenommen werden.

tifier nicht vorherrschen. Es gibt im Stück jedoch die Rolle “mac-all”. Diese ist sicherlich nicht von allen im Stück vorkommenden Charakteren zu sprechen, sondern bezieht sich auf die Personen in der Szene. So entstehen semantische Fragmente, die inhaltlich aufbereitet werden können.
Dabei zeigt sich hier eine Auszeichnungspraxis, die nicht sauber ist. Dies kann in verschiedenen Auszeichnungsinterpretationen anders gehandhabt werden. Damit würden Daten im gleichen Format nicht gleich behandelbar. Die Vergabe der IDs würde des Weiteren eine wirkliche Vergleichbarkeit von Werksinterpretationen erschweren, da die IDs in verschiedenen Dokumenten einander zugeordnet werden müssten.
Im vergleichbare extraktionen und Visualisierungen von Netzen ermöglichen es Stücke zu Vergleichen und Muster zu finden. Bei Untersuchungen zu Rollen und sozialen Konstrukten konnten wenige wiederkehrende Muster für griechische Tragödien gefunden werden. Die Stücke ließen sich daraufhin nach ihrem sozialen Netzwerk klassifizieren.[214]

Text

Da es sich beim Stück primär um eine Textquelle handelt, wäre es wichtig zu wissen, was im Text steht und wie es in Beziehung zu Szenen und Charakteren steht. Hierzu stehen Textextraktionsmethoden bereit.
Hier können Worte sowie N-Gramme[11] zur bestehenden Visualisierung hinzugefügt werden. Hier stellt sich die Frage, welche Textfragmente sind relevant genug, um angezeigt zu werden. Das muss durch einen Filter bestimmt werden.
Kurze, aber aussagekräftige, mindestens zweimal, vorkommende Textfragmente wären wünschenswert. Zu lange Textfragmente lassen sich jedoch nicht ohne Überschneidungen als beschriftete Knoten darstellen. Daher sind einzelne Worte, Bi- oder Trigramme interessant.
Hier sollten möglichst wenig Textfragmente ausgewählt werden, die möglichst aussagekräftig sind. Zu viele Knoten würden die Darstellung überladen. Vor allem Füllworte, Kon- und Disjunktionen und Artikel enthalten wenig spezifische Information, kommen jedoch häufig vor.
Verbindende Wortfragmente, die nicht als Attribute von Knoten, sondern Verbindungen zwischen Knoten auftreten, sind wünschenswert. Textfragmente, die nur für Stücke oder Einzelrollen wichtig sind, helfen bei der Beschreibung eines Charakters oder einer Szene, sie helfen jedoch nicht beim Verständnis des Kontexts oder der thematischen Zusammenfassung.
Gegen das Vorkommen von N-Grammen spricht dabei die künstlerische Gestaltung des Werks. Wiederholungen von Wortabfolgen werden daher eher vermieden. Es wurden im Text aber auch keine signifikanten Bi- oder Trigramme gefunden. Daher wird mit Tokens aus Einzelworten gearbeitet.
Im weiteren wird eine einfaches TF-IDF-Maß[215] verwendet, um die einzelnen Charaktere und Szenen mit Worten zu verknüpfen. Bei diesem Verfahren werden die Termfrequenz

[11] Mehrwortkombinationen, engl.: n-gram

(TF) und inverse Dokumentfrequenz (IDF) ins Verhältnis gesetzt. Die Termfrequenz gibt an, mit welcher Häufigkeit ein Wort von einem Charakter im gesamten Stück oder in von allen Charakteren einer Szene gesprochen wird. Die Dokumentfrequenz (DF) besagt, wie häufig ein Wort insgesamt über alle Szenen und Charaktere verwendet wird. Invertiert (IDF) heißt das, wie speziell ein Wort verwendet wird. Bei diesem Verfahren werden charakteristische Tokens für Charaktere und Szenen aus deren Text ermittelt.
Charaktere und Szenen haben ihren eigenen Corpus, den aus dem Text aller Charaktere und den aller Szenen. Dabei ist die Gliederung für das Errechnen der Werte signifikant. Ein hoher Wert besagt, dass ein Wort häufig im Dokument und speziell in diesem Dokument vorkommt. Eine gewichtete Kante zwischen Charakter und Token oder Szene und Token. Die Auswahl der gewichteten Tokens wird im nächsten Schritt vorgenommen. Das relevanteste Prozent aller Worte der Charaktere und Szenen wird zusammengefasst. Diese Worte werden dann in die bestehende Visualisierung eingefügt. Die Kanten, deren Worte nicht erfasst sind, verfallen.
Die Kombination der Begriffe ist in Darstellung 3.5[12] weiter erläutert.

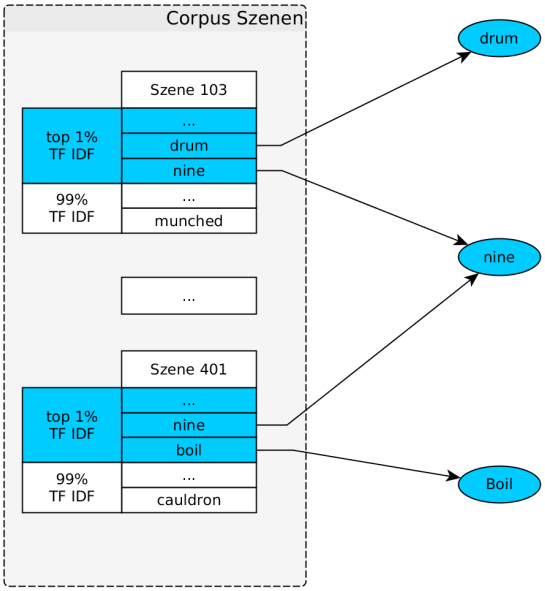
Zu beachten ist, dass nicht alle entstandenen Kanten signifikante TF/IDF Werte repräsentieren. Diese Kanten entstehen aus der Auswahl eines Begriffs als signifikanten Begriffs für eine Szene oder eine Rolle. Diese Verbindungen werden in Abbildung 3.5 als gestrichelte Linien dargestellt. Durch das beschriebene Vorgehen kommen Worte als neuer Knotentyp in die Visualisierung. Diese werden abhängig von der vorherigen Grundordnung durch kräftebasierte Positionierungsverfahren positioniert. Zur Erhaltung der Grundordnung werden die Szenen und Charaktere fixiert. Die Knoten werden so vom Positionierungsalgorithmus nicht verschoben. Die Wortknoten ordnen sich zwischen den fixierten Knoten an.
Um lokal wichtige Worte hervorzuheben, wurde für die Größe der Labels der Clusterkoeffizient als Gewichtung gewählt. Auf diese Weise werden Begriffe hervorgehoben, wenn sie in einem engen Kontext von Personen in den Szenen stehen. Das ist ein erwünschter Nebeneffekt der Verbindung von Szenen mit Personen, denn nur so können aus dem vormals bipartiten Netz Dreiecke entstehen, die vom Clusterkoeffizienten herangezogen werden.
Abbildung 3.6 zeigt das Ergebnis dieser Visualisierung. Die so entstandene Visualisierung lässt sich nicht einfach ohne Zoomen lesen. Abbildung 3.7 zeigt einen Ausschnitt mit den Hexen. Hier sind spezielle Hexenworte durch die Gewichtung hervorgehoben, da nur die Hexen und die Szenen mit den Hexen diese Worte beinhalten. Wortfragmente oder Wortabfolgen haben den Vorteil, dass sie automatisch erkannt werden können. Jedoch bleiben so einzigartige Phrasen verborgen. Diese stellen wahrscheinlich keine signifikanten Muster dar, da sie im Zweifel nur in einer Szene von einer Rolle gesprochen werden.
Ein anderer Ansatz wäre die Vorauswahl von Zitaten oder herausstechenden Phrasen. Diese würden mit Szenen und Charakteren verbunden und eine Einordnung einer berühmten Phrase in den Kontext des Stücks auf visuelle Weise ermöglichen. An dieser Stelle tritt

[12]Die Grafik ist an den Daten angelehnt. Sie repräsentiert jedoch nur einen Ausschnitt und hat einen beispielhaften Charakter. Die Grafik soll das Vorgehen und die Kombinationen des Verfahrens verdeutlichen. Daher können Abweichungen zu den Quelldaten entstanden sein.

ein Auswahlproblem auf. Hier würde die Auswahl der Zitate entscheidend sein, da je nach Länge des Texts nur eine sehr begrenzte Anzahl von Zitaten eingefügt werden kann. Dieses Auswahlproblem kann nicht algorithmisch durch den reinen Text gelöst werden. Hier wäre mehr Kontext, eine spezielle Datenquelle oder ein Experte aus Fleisch und Blut vonnöten.

Abbildung 3.5: *Beispielhafte Darstellung des TF/IDF Verfahrens zur Auswahl von entitätenspezifischen Worten.*

(a) *Auswahl von szenenspezifischen Werten.*



(b) *Kombination der rollenspezifischen mit den szenenspezifischen Worten.*

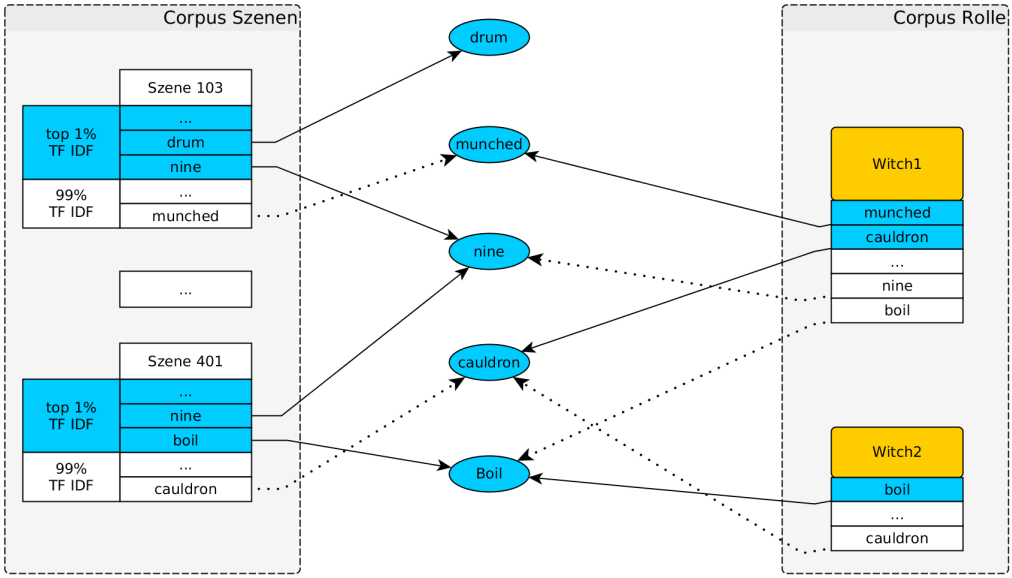


Abbildung 3.6: *Visualisierung aller Szenen und Rollen in Macbeth mit den Worten aus dem TF/IDF verfahren.*

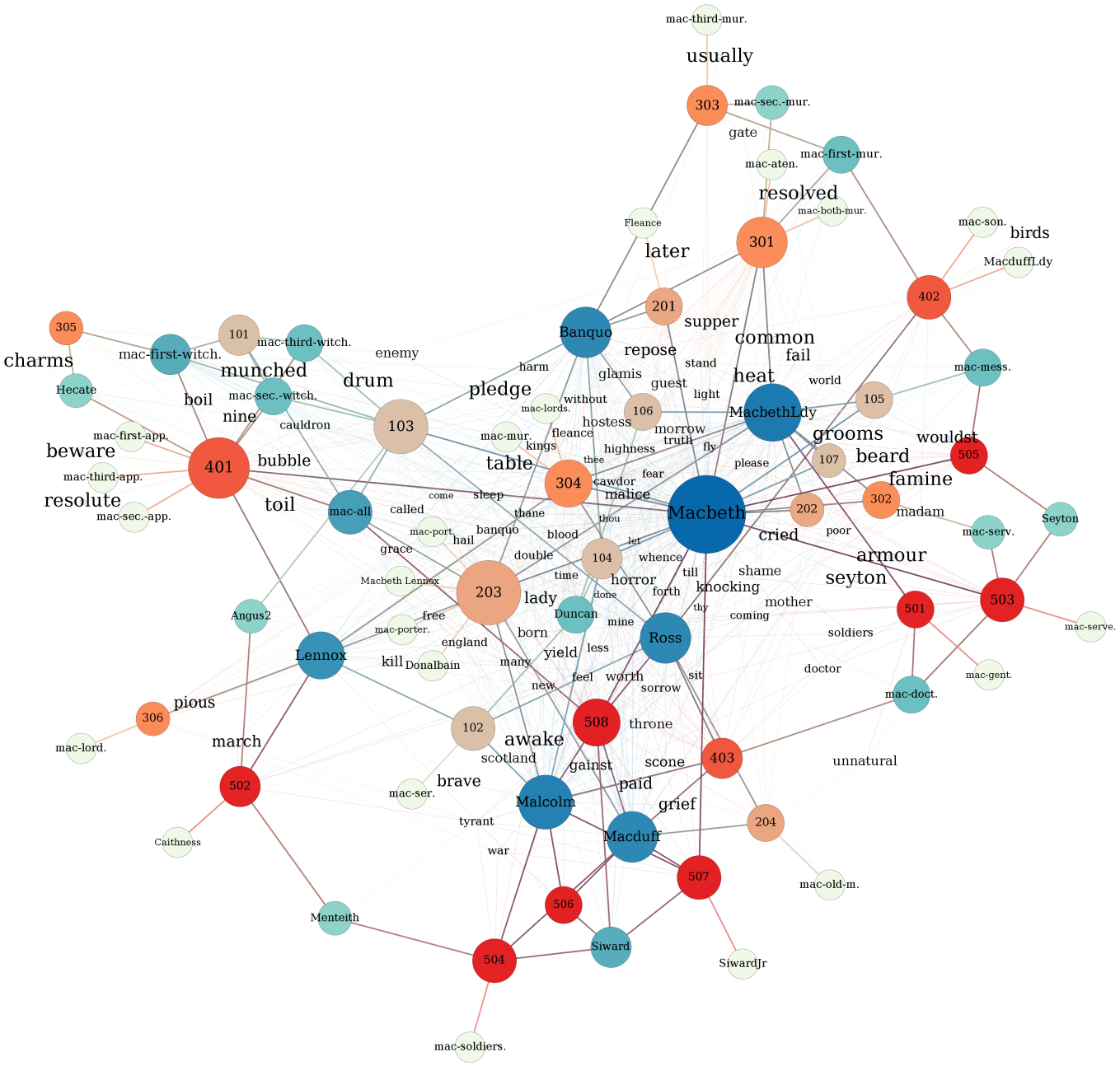
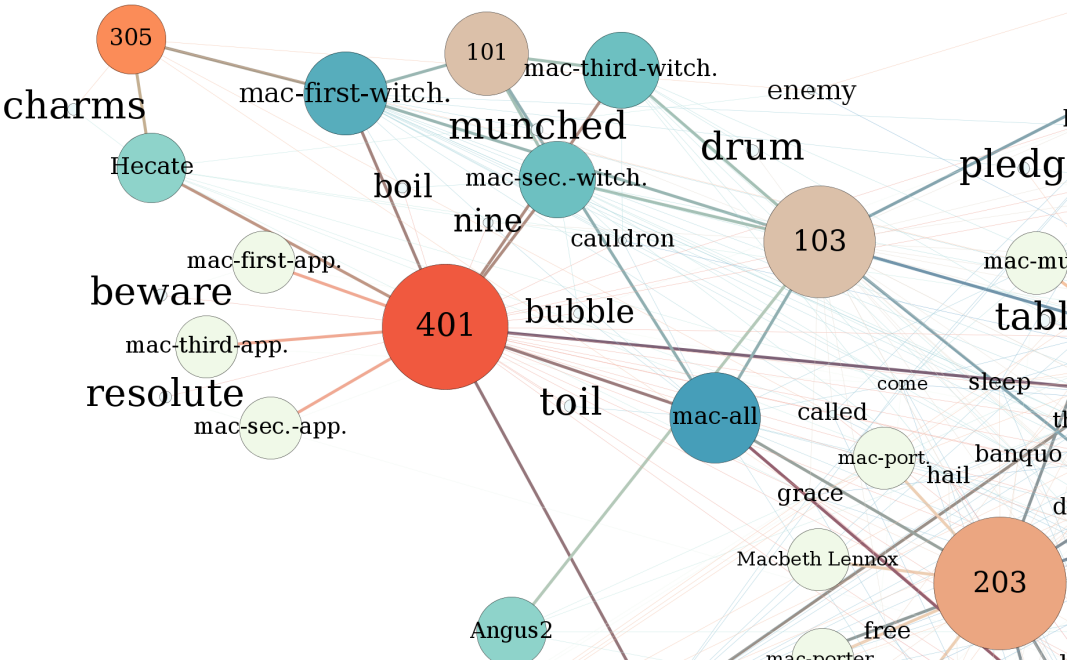


Abbildung 3.7: *Detail aus 3.6 mit Fokus auf den Hexen.*



3.2.2 Gesetze

Ein weiteres Beispiel für die Analyse von XML in TEI sind Gesetzestexte. Das Bundesministerium der Justiz bietet Gesetzestexte in XML auf seiner Homepage[103] an. Hierzu wird eine eigene DTD als Beschreibung der Datenstruktur bereitgestellt.
Im weiteren wird das Bürgerliche Gesetzbuch (BGB) auf seine Struktur untersucht. Das BGB folgt der Klammertechnik, die sich vom Ausklammern in der Mathematik ableitet. Das heißt: Faktoren, die für eine Menge von Gesetzen gleich sind, lassen sich vor diese Gesetze ziehen und werden damit vor ihrem Gebrauch definiert. Konkret gibt es im BGB allgemeine und spezielle Teile, wobei sich die speziellen Teile auf die allgemeinen Teile beziehen. Dadurch wird an den meisten Stellen zurückverwiesen und nicht nach vorne.[272, 13f] Diese Regeln werden jedoch durch die stetige Arbeit an den Gesetzestexten durchbrochen und es kommt zu Vorverweisen.
Im Folgenden wird gezeigt, wie die Auszeichnung eines Textes als Grundlage für die Extraktion eines Netzes verwendet werden kann. Weiterhin wird gezeigt, wie mit einem iterativen Vorgehen weitere Nennungen von Elementen erkannt und genutzt werden können die im Ausgangstext nicht explizit markiert sind.

Struktur und Extraktion

In den Gesetzestexten im XML-Format ist die Struktur der Gesetzbücher hinterlegt. Dabei können Gesetze mit ihren Paragrafen einfach identifiziert werden. Jedes Gesetz und jeder Paragraf ist ausgezeichnet und in einem Attribut nach einem einheitlichen Schema benannt.
Eine granularrere Abbildung und Strukturierung in Absätze von Paragrafen ist dabei auch möglich.[222] Die Gesetze sind im Gesetzestext referenziert. Zitate auf andere Gesetze sind jedoch nicht explizit mit Referenz, Tags oder Attributen ausgezeichnet. Die Gesetzesreferenzen liegen in menschenlesbarer Form als Freitext vor.
Jedoch bietet das Paragrafenzeichen "§" einen Hinweis auf ein Gesetzeszitat. Um einen Referenzgraphen zwischen Gesetzestexten zu finden, ist es nötig, die XML-Dateien durch Gesetzeszitate zu erweitern. Dafür sind wiederum Verfahren nötig, um Gesetze in Texten zu erkennen, also eine Unterklasse von Freitext zu analysieren können. Das kann durch die Angabe von Paragrafen erledigt werden. Diese beziehen sich jedoch nicht zwangsläufig auf das gleiche Gesetzbuch, sondern auch auf andere Gesetzbücher. Gesetzeszitate liegen vielleicht in für den Menschen eindeutiger, aber nicht in maschinenlesbarer Form vor. Somit sind sie nicht einfach maschinell verarbeitbar. Vorgehen diese Daten maschinell zu erfassen, gehen wahrscheinlich mit Fehlern einher.
Beim Parsing traten folgende Probleme auf:

1. Gesetzeszitate können sich nicht nur auf einzelne Paragrafen beziehen, sondern auf Reihen von Paragrafen mit Ausnahmen.

2. Zitate können sich auch auf einzelne Sätze in Paragrafen beziehen.

3. Es kann sich auf Paragrafen anderer Gesetzbücher bezogen werden.

4. Referenzen, die sich auf explizite Nichtanwendung beziehen, Negativzitate.

5. Die Komplexität natürlicher Sprache. Zitate können beliebig komplex sein.

Im Weiteren wird die XML-Datei zum Bürgerlichen Gesetzbuch (BGB) betrachtet. Es wurde, mit einem einfachem Parser auf Grundlage von regulären Ausdrücken (RegExp)[13] aus dem Gesetzestext ein Netzwerk aus Paragrafen erstellt.
Die RegExp gliedern die Paragrafenzitate in ein paar einfache Muster. Zitate werden grundlegend in zwei Klassen eingeteilt: Einfachzitate mit einem Paragrafen Zeichen "§" und Mehrfachzitate mit doppeltem Paragrafen "§§". Dazu kommen reguläre Ausdrücke, die Verweise auf andere Gesetzbücher erkennen, damit diese herausgefiltert werden können. Auch mussten bei der Untersuchung immer wieder händisch die Referenzen überprüft werden, um so eine Vielzahl von Besonderheiten wie Flexionen von Signalwörtern in den RegExp-Mustern zu erfassen.
Ein Beispiel für eine solche Fehlerquelle ist die Verwechselung von Absatzangaben mit Paragrafen. Hier müssen sprachliche Flexionen wie "Absatzes", "Absätzen" etc. separat erfasst werden.
Ein anderer auffälliger Fehler ergibt sich aus den die Relationen zu den ersten Paragrafen, vor allem 1 bis 9. Hier zeigen sich Fehler darin, dass Angaben wie Verweise auf Sätze als Paragrafen interpretiert wurden oder Zahlen von regulären Ausdrücken nur teilweise erfasst wurden. Dies führte zu Knoten mit einem sehr hohem Grad. Sehr hohe Abweichungen im Grad wiesen immer wieder auf Muster im Text hin, die speziell gehandhabt werden mussten.

Untersuchung

Im Anschluss an die Konstruktion eines sauberen Graphen können nun einfache strukturelle Fragen gestellt werden:

1. Wie verteilen sich die Referenzen?

2. Greifen Paragrafen in ihrer Definition auf Paragrafen vor? Refenziert ein Paragraf einen anderen, der noch definiert wird?

3. Sind Referenzen eher lokal, verweisen also auf nahe Textstellen oder sind sie global und verweisen auf entfernte Teile des Buchs?

Im Folgenden werden die Fragen am Beispiel der größten Komponente beantwortet. Unverbundene Paragrafen, Paragrafen, die nicht referenziert wurden oder nicht referenzieren, wurden für die Analyse entfernt.
Bei der Untersuchung fällt auf, dass einige wenige Paragrafen sehr häufig zitiert werden. § 210[14] wurde am häufigsten, 13 mal, zitiert. Er befasst sich mit nicht voll geschäftsfähigen

[13] Englisch: Regular Expressions oder kurz RegExp.

[14] Der vollständige Text ist unter `http://www.gesetze-im-internet.de/bgb/__210.html` verfügbar. Zuletzt geöffnet 02.01.2015

Abbildung 3.8: *Zirkeldarstellung des Zitatnetzes. Die Paragrafen sind im Uhrzeigersinn sortiert, rote Kanten sind Rückverweise. Blaue Kanten sind voraus Verweise. Die Dicke der Kanten zeigt die Paragrafendistanz der Verweise an.*

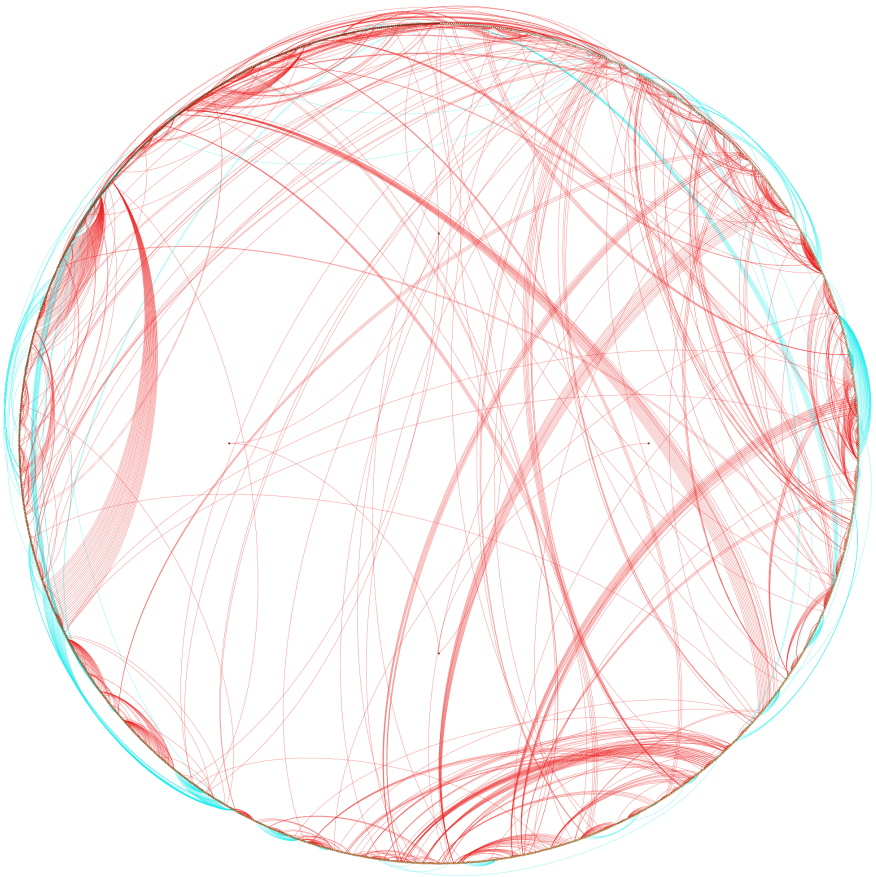


Tabelle 3.1: *Top 10 Paragrafen BGB nach ausgehendem Grad*

Ausgehender Graf	Paragraf	Titel
90	549	Auf Wohnraummietverhältnisse anwendbare Vorschriften
62	1908i	Entsprechend anwendbare Vorschriften
60	1318	Folgen der Aufhebung
34	1518	Zwingendes Recht
28	1458	Vormundschaft über einen Ehegatten
25	1933	Ausschluss des Ehegattenerbrechts
24	1273	Gesetzlicher Inhalt des Pfandrechts an Rechten
23	512	Anwendung auf Existenzgründer
21	511	Abweichende Vereinbarungen
20	578	Mietverhältnisse über Grundstücke und Räume

Abbildung 3.9: *Darstellung des BGB mit einem kräftebasierten Positionierungsverfahren in Gephi. Der Gradient auf den Knoten gibt die Position im BGB an. Weiße Knoten sind Paragrafen am Anfang, orange in der Mitte, Braun am Ende, ansonsten wie Abbildung 3.8.*

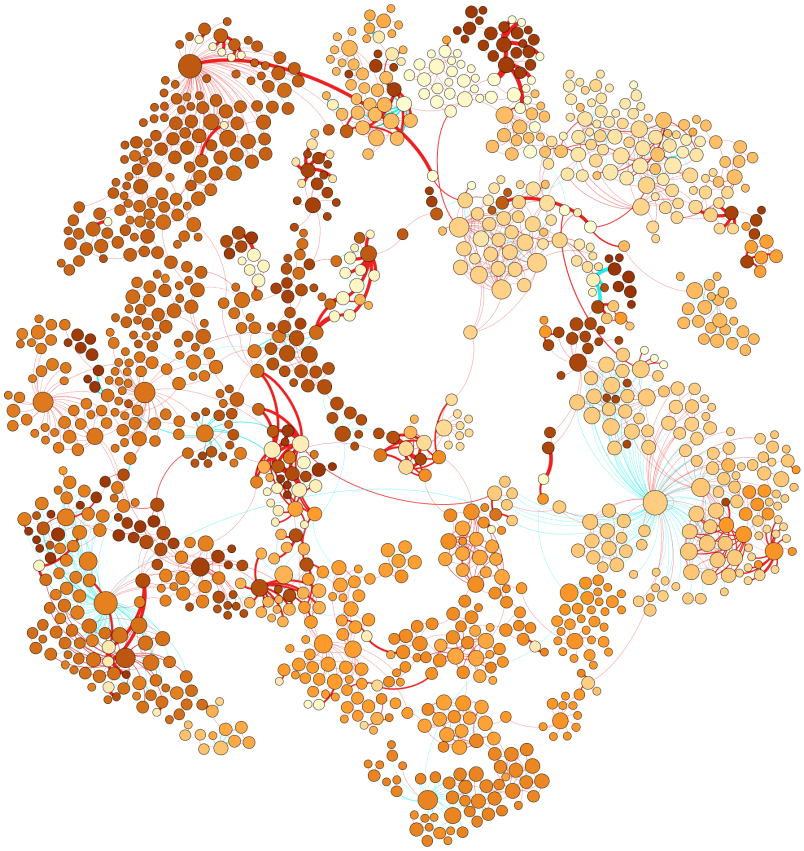
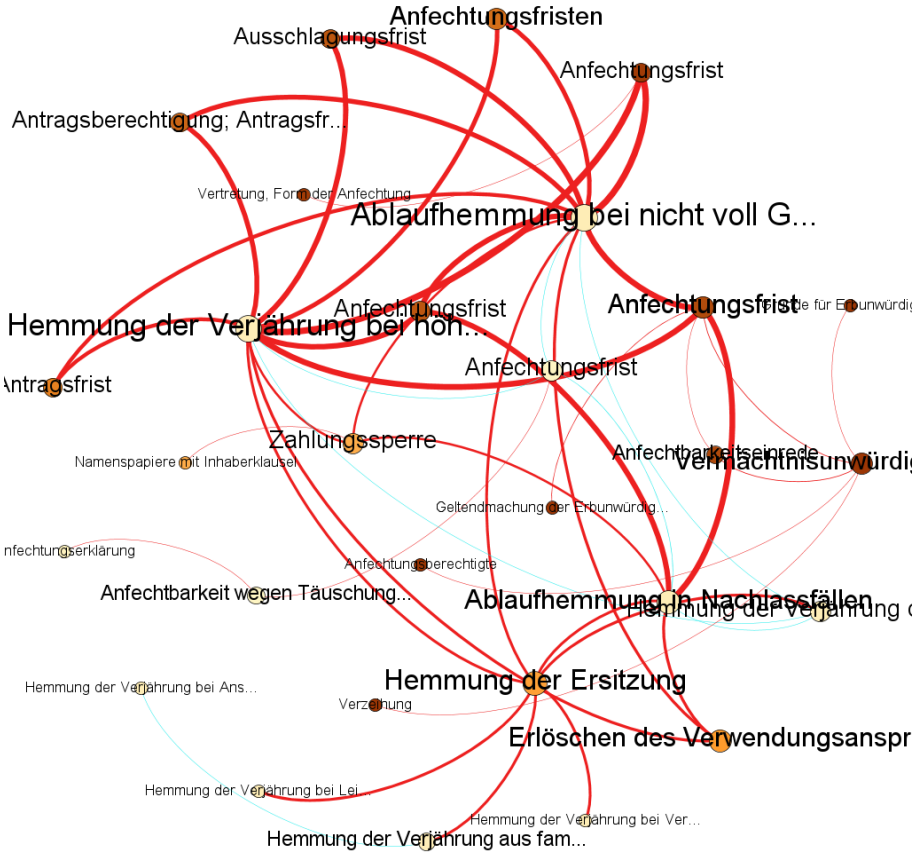


Tabelle 3.2: *Top 11 Paragrafen BGB nach eingehendem Grad*

Eingehender Grad	Paragraf	Titel
13	210	Ablaufhemmung bei nicht voll Geschäftsfähigen
12	206	Hemmung der Verjährung bei höherer Gewalt
10	559	Mieterhöhung nach Modernisierungsmaßnahmen
9	566	Kauf bricht nicht Miete
9	566a	Mietsicherheit
9	566c	Vereinbarung zwischen Mieter und Vermieter über die Miete
9	566d	Aufrechnung durch den Mieter
9	566e	Mitteilung des Eigentumsübergangs durch den Vermieter
9	873	Erwerb durch Einigung und Eintragung
9	876	Aufhebung eines belasteten Rechts
9	1807	Art der Anlegung

Abbildung 3.10: *Ausschnitt aus 3.9 mit den Titeln der Paragrafen als Beschriftung. Layout erfolgte mit dem Fruchtermann-Rheingold Algorithmus[99] mit Gephi.*



Menschen und Verjährungsfristen. Auch § 206, an zweiter Stelle, befasst sich mit Verjährungsfristen. Danach gibt es eine Gruppe von Paragrafen, die sich auf das Mietverhältnis nach Verkauf eines Hauses beziehen. Sie stehen alle hintereinander und haben den gleichen Grad. Das lässt vermuten, dass sie auch im gleichen Kontext zitiert wurden. Die Liste der 11[15] am häufigsten zitierten Paragrafen findet sich in Tabelle 3.2.
Es gibt sehr viel mehr Paragrafen, die viele andere Paragrafen zitieren. Darunter sind Listen von anwendbaren Vorschriften, beispielsweise § 549 "Auf Wohnraummietverhältnisse anwendbare Vorschriften"[16]. Diese Paragrafen bilden eine Abgrenzung für einen Bereich anwendbarer Regeln zu einem Thema.
Einer der Top 10 Paragrafen findet sich in der Auflistung 3.1.
Abbildung 3.8 zeigt die Anordnung der Paragrafen in ihrer logischen Reihenfolge. Die Abfolge läuft dabei von § 1, oben am Kreis, im Uhrzeigersinn. Die Dicke der Kanten gibt einen Eindruck von der "Distanz" in der "Einheit" Paragrafen, die diese Kanten überwinden. Die Farbe besagt, ob sie Vor- oder Rückwertsreferenzen sind. Verweise auf Paragrafen mit höherem Wert als der zitierende Paragraf sind Blau und Verweise auf schon genannte Paragrafen sind Rot. Dabei ist zu sehen, dass die meisten "weiten" Verweise zurückverweisen. Auch im Ganzen gibt es mehr zurückverweisende Referenzen. Diese Beobachtung war bei der Anwendung der Klammertechnik zu erwarten.
Abbildung 3.9 ging aus der Vorsortierung von Abbildung 3.8 hervor und wurde mit einem Force-based Algorithus erstellt in dessen Einstellungen die Hubs nach außen getrieben wurden. Dies sollte helfen ein Cluster von zentralen Knoten zu bekommen, da dies auch alle anderen Knoten in die Mitte drängt. Die Knoten, die in Abbildung 3.8 nicht zu sehen waren, sind hier nun erkennbar. Ihre Farbintensität gibt ihre Position an. § 1 ist ein fast weißer Knoten, § 1000 ist orange und der letzte Paragraf (§ 2370) ist dunkelbraun. Diese Farbwahl hilft, wie bei Macbeth (3.2.1), die Linearität in der Anordnung zu erkennen. Die zirkuläre Anfangsanordnung ist in dieser Darstellung noch dominant. Doch gibt es auch vereinzelt Cluster mit verschiedenfarbigen Knoten.
Eines dieser Fragmente ist in Abbildung 3.10 zu sehen. Hier sieht man auch eine viele Knoten, die aus der Beschreibung der Gradzahlen bekannt sind. In diesem Cluster von Paragrafen geht es um Verjährungsfristen, Hemmungen und andere formale Kriterien. Diese scheinen textübergreifend von Interesse zu sein. Daher scheint der lineare Charakter des Textes an dieser Stelle am meisten gebrochen zu werden.

Fazit

Die Darstellungen in diesem Abschnitt haben die lineare und die nichtlineare Struktur eines Gesetzestextes aufgezeigt. Dazu wurde ein iteratives Vorgehen zum Extrahieren der Zitate aus dem Text gewählt. Die Extraktion der einzelnen Gesetzeszitate wurde mit einer Menge von regulären Ausdrücken bewerkstelligt. Die stetige Visualisierung während

[15] Es sind elf Ergebnisse, da mehrere Knoten einen eingehenden Grad von 9 haben und damit gleich bedeutend sind.

[16] Für vollständigen Text siehe `http://www.gesetze-im-internet.de/bgb/__549.html` zuletzt geöffnet am 02.01.2015.

der Entwicklung der regulären Ausdrücke hat eine entscheidende Rolle gespielt, denn so wurden Fehler frühzeitig sichtbar.
In dieser Untersuchung musste die Granularität der Referenzen gewählt werden. Dazu wurden Referenzen immer auf Paragrafenebene extrahiert. Die Referenzen hätten genauer betrachtet werden können. Dazu hätten die Referenzen auf Abschnitte und Nebensätze von Paragrafen mit berücksichtigt werden müssen.
Eine weitere Untersuchung könnte die Verbindungen der Gesetzestexte untereinander betrachten. Dieses Vorgehen würde andere Sachverhalte wie die Interdependenz dieser Texte beleuchten.
Auch bietet die Strukturierung des Textes Möglichkeiten, Wort- und Textanalysen durchzuführen, wie bei Schönhof et al. vorgeschlagen.[222]
Da die Daten Links auf die Volltexte der Paragrafen bereitstellen, wäre es ohne Weiteres möglich, eine interaktive Grafik aus diesen Daten zu erstellen. Die einmal erschlossene Struktur ermöglicht es, auch andere Gesetzestexte auf diese Weise zu untersuchen. Die hier gezeigten Untersuchungen zur Linearität eines Gesetzestextes können auf andere Texte im gleichen Format angewendet werden. Dadurch könnte die Komplexität der einzelnen Texte verglichen werden und ob auch diese der Klammertechnik folgen.

3.3 Stimmbezirke

Dieser Abschnitt beschäftigt sich mit der Auswertung und Transformation von Tabellen und Geodaten im Rahmen der Graphenanalyse. In diesem Abschnitt wird mit kontinuierlichen Daten gearbeitet. Sie haben den Vorteil, dass sie sich sehr viel besser objektiv bewerten lassen als kategoriale Daten, wie sie in den XML-Beispielen verwendet wurden. Anhand der Ergebnisse der Bundestagswahl aus dem Jahr 2013 wird gezeigt, wie kontinuierlichen Daten in Ähnlichkeitsnetze umgewandelt werden können. Dabei werden die Möglichkeiten von geografischen und kräftebasierten Anordnung gegenübergestellt. Hierzu wird die geografische Einteilung des Wahlverhaltens auf Städteebene verwendet. Clustering-Algorithmen und Farben werden verwendet, um eine Referenz zwischen den beiden Darstellungsweisen zu ermöglichen.

3.3.1 Untersuchung

Datengrundlage

Als Datengrundlage dienen die Wahlstrukturdaten der Bundestagswahl 2013 der Stadt Köln und der Stadt Stuttgart. Dabei sind die Wahldaten nach Stimmbezirken aufgeschlüsselt. In den Daten ist die Information enthalten, wie viele Menschen in welchem Stimmbezirk welcher Partei ihre Zweitstimme gegeben haben. Die Wahlergebnisse liegen dabei als Liste vor.
Weitere Informationen sind die Geo-Daten der Stimmbezirke als georeferenzierte Vektorgraphiken. Diese Daten werden als Open-Data von Köln[45, 227] und Stuttgart[46] bereitgestellt. Die Daten sind sehr ähnlich strukturiert und inhaltlich vergleichbar.

Ziel der Darstellung

Eine netzartige Betrachtung ist bei diesen Daten nicht offensichtlich. Wahldaten sind keine klassisch relationalen Daten. Es sind Daten, die normalerweise mit statistischen Methoden analysiert werden. Diese Daten sagen aus, wie viele Menschen in einem klar beschriebenen Gebiet welche Partei wählten.
Die Betrachtung der Daten bezieht sich inhaltlich nicht direkt auf den Ausgang der Bundestagswahl, sondern soll eine mikro-geografische Verteilung der politischen Präferenzen aufzeigen.
Die Aufaddierung der Einzelergebnisse beeinflusst zwar die Zusammensetzung der Regierung aber im Gesamtergebnis verschwindet die interne Mikrostruktur der Stadt.
Für die Untersuchtung wird die Zweitstimmenpräferenz verwendet. Diese wird als vergleichbarer angenommen als die Erststimmenverteilung. Erststimmen sind kandidatengebunden und können daher je nach Kandidat und lokaler Agenda in einem Stimmbezirk eher variieren als die Zweitstimmen.
Ziel ist es eine mikro-geographische Darstellung, die an der Parteienlandschaft in Deutschland orientiert ist. Dies lässt Rückschlüsse auf die Verteilung politischer Meinungen und der Sozioökonomischen Verteilung innerhalb von Städten zu.

Vorgehen

In dieser Betrachtung sind 7 Parteien enthalten: CDU, SPD, Grüne, Linke, FDP, AfD und Piraten. Eine Betrachtung, wie sie im Weiteren vorgenommen wird, ist mit wenig Items wenig sinnvoll. Simplere und besser erprobte Ansätze sind vielversprechender. Nachfolgend wird einfache deskriptive Statistik verwendet. Hierbei ist der Vergleichsrahmen die Stadt selbst und nicht die Stadt in Bezug auf das bundesweite Gesamtergebnis der Wahl.
Im ersten Schritt wird für alle Stimmbezirke eine positive oder negative Präferenz für eine Partei auf Stimmbezirksebene errechnet.
Eine Präferenz eines Stimmbezirks für eine Partei kann man daran erkennen, ob die Partei über- oder unterdurchschnittlich abgeschnitten hat. Hier wird vernachlässigt, dass sich die Größe der Stimmbezirke, die Wahlberechtigten in diesem Stimmbezirk und die Wahlbeteiligung im Stimmbezirk mitunter doch unterscheiden. Diese Verzerrungen werden vernachlässigt, da Stimmbezirke eigentlich recht zuverlässig zugeschnitten sind und etwa 1.000 Einwohner umfassen.
Einige Rechenbeispiele:
Partei X hat in der gesamten Stadt A 20% der Stimmen.
In Stimmbezirk A1 ist der Anteil bei nur 5%.
In Stimmbezirk A2 hat Partei X 60% Stimmenanteil.
In Stimmbezirk A3 hat Partei X 22% Stimmenanteil.
In Stimmbezirk A1 hätte Partei X ein “schlechtes” Wahlergebnissen in Stimmbezirk A2 ein “sehr gutes” Wahlergebnis und in Stimmbezirk A2 ein durchschnittliches Ergebnis.
Partei X hat in Stadt B 10 % der Stimmen, in Stimmbezirk B1 5% der Stimmen und
In Stimmbezirk B2 20 % der Stimmen.
In Stimmbezirk B1 ist Partei X unterdurchschnittlich, es zeigt sich aber keine so große

Abweichung wie in Stadt A, Stimmbezirk B2 ist aber selbst mit 20% überdurchschnittlich im Verhältnis zum städteweiten Ergebnis. Darin ähnelt Stimmbezirk B2 dem Stimmbezirk A1.
Für eine Analyse muss eine Grenze für das "überdurchschnittliche" oder "gute" Ergebnis gezogen werden. Anderenfalls würde das erzeugte Ähnlichkeitsnetz sehr dicht werden. Dadurch hätten einzelne Kanten wenig Aussagekraft. Es wird ein Schwellwert benötigt, der überschritten werden muss, um eine Ähnlichkeit oder Präferenz zu codieren. Hier wurde aus pragmatischen Gründen die Standardabweichung gewählt.[17]
Weicht in einem Stimmbezirk die Anzahl der Stimmen für eine Partei vom Durchschnitt der Stimmen in der Stadt um die Standardabweichung ab, dann wird angenommen, dass dieser Stimmbezirk eine Präferenz für diese Partei hat. Diese Präferenz kann positiv und negativ ausfallen, je nach der Abweichung über oder unter den Schwellenwert der Standardabweichung.
In Abbildung 3.11 sind die beschriebenen Wahlausreißer zu sehen. Dabei werden alle Stimmbezirke unterhalb der roten Linie dem negativen Item einer Partei zugeordnet, während alle Bezirke oberhalb der grünen Linie dem positiven Item zugeordnet werden.
Neu codiert bedeutet diese Abweichung eine Verbindung zu einer Partei. Dieser Schritt verwandelt die Wahldaten von einer kontinuierlichen, metrischen Skala in nominale Items. Partei X kann also in Stimmbezirk A1 eine negative Präferenz verzeichnen, während sie in Stimmbezirk A2 eine positive Präferenz verzeichnen kann. Stimmbezirk A3 mit einem Wert von 22% wird nach dieser Rechnung weder mit einer negativen noch einer positiven Tendenz gesehen, da sich die Zustimmung innerhalb der Standardabweichung befindet.
Daraus ergibt sich ein bipartites Netz. Die Knoten der ersten Seite sind bilden die Stimmbezirke ab und die Knoten der zweiten Seite die Parteipräferenzen. Die Parteipräferenzen spalten sich dabei in negative und positive Präferenzen. Daraus ergeben sich bei 7 Parteien 14 mögliche Präferenzen zu Parteien. Diese können sich überschneiden, außer natürlich Präferenzen eines Stimmbezirks zur negativen und positiven Präferenz der gleichen Partei. Die Zuordnung eines Stimmbezirks zu beiden Items negativ CDU und positiv CDU ist daher nicht möglich.
Eine Projektion des bipartiten Netzes auf die Seite der Stimmbezirke lässt ein Netz von Ähnlichkeiten der Stimmbezirke entstehen. Dieses Netz zeigt bei der Visualisierung eine Aufteilung in politische Lager, die sich über positive und negative Präferenzen von Parteien definiert. Die Projektion soll eine bessere Verteilung der Bezirke und ihrer Präferenzen zueinander zu ermöglichen.
Die Visualisierung zeigt dabei, dass sich von der Präferenz ähnliche Stimmbezirke näher sind. Da es sich bei den Stimmbezirken um eine geografische Information handelt, ist es möglich, die Information aus dem Netz auch geografisch abzubilden.
Dafür wurden die Knoten des Netzes aufgrund ihrer Ähnlichkeitskanten geclustert. Das

[17]Das birgt zwar immer noch eine gewisse Willkür, denn das Überschreiten der Toleranzgrenze um ein Hundertstel wird genauso kategorisiert wie das Überschreiten um mehrere Prozent. Dieser Konflikt lässt sich nicht endgültig lösen, wenn man eine binäre Zuordnung von Gruppen verfolgt. Ein Ansatz, der hier Abhilfe schaffen könnte, wären beispielsweise fuzzy sets.

dafür genutzte hierarchische Clusteringverfahren wurde so parametrisiert, dass vier Cluster entstehen. Die Wahl der Menge an Clustern ist bei Hierarchischen Clusteralgorithmen notwendig. Die Cluster haben dabei keine explizite Bedeutung ihre Zusammenstellung geht auf die Struktur des Ähnlichkeitsnetzes zurück. Durch das Clustering wird in erster linine die Komplexität reduziert.
Es entstehen dabei nicht vollständig klar benennbare Cluster, Abbildung 3.12a zeigt die Namen der Präferenzen und somit eine etwaige Position der Präferenzen im unipartiten Graphen.[18]
Grob lassen sich die Cluster für Köln wie folgt benennen:

Grau bildet einen Block aus CDU und FDP sowie einer ablehnenden Position von Piraten, Linken und SPD.

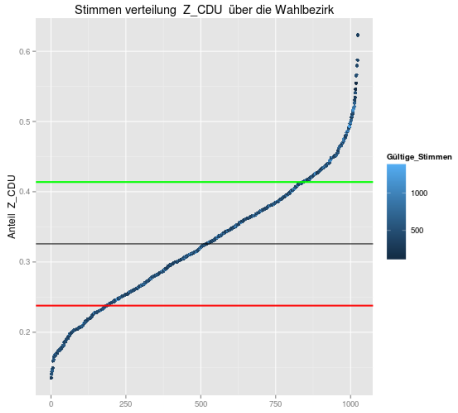
Orange hängt mit der AfD und einer Ablehnung der Grünen zusammen. Ansonsten gibt es eine positive Präferenz zur SPD.

Rot ist eher bei Piraten, SPD und Linken zu verorten und hängt mit einer Ablehnung der FDP zusammen.

Lila zeichnet sich durch Zustimmung zu den Grünen und der Ablehnung der CDU und AfD aus, mit Nähe zur Linken.

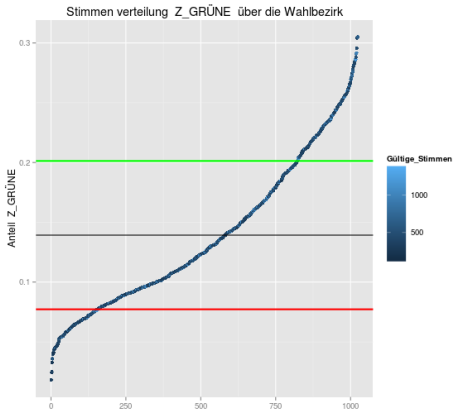
[18]Die Positionierung der Stimmbezirke erfolge über die Kanten der Projektion auf die Stimmbezirke über die Präferenzen. Nach dieser Positionierung wurden die Knoten fixiert. Die Kanten des bipartiten Netzes wurden danach benutzt, um die Bezeichnung der Präferenzen zu positionieren.

Abbildung 3.11: *Die Verteilung der Stimmen nach Stimmbezirken und Parteien für Köln. Die Stimmbezirke wurden nach ihren Stimmen für eine Partei sortiert. Alle Stimmbezirke oberhalb der Grünen Linie erhalten eine positive Parteipräferenz. Alle Stimmbezirke unterhalb der roten Linie erhalten eine negative Parteipräferenz.*

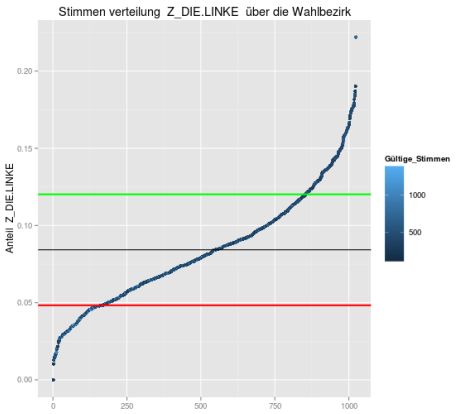


(a) CDU

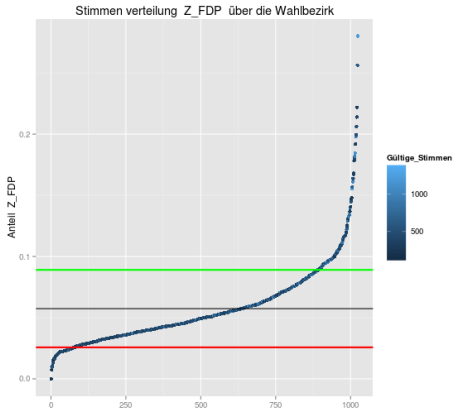
(b) SPD



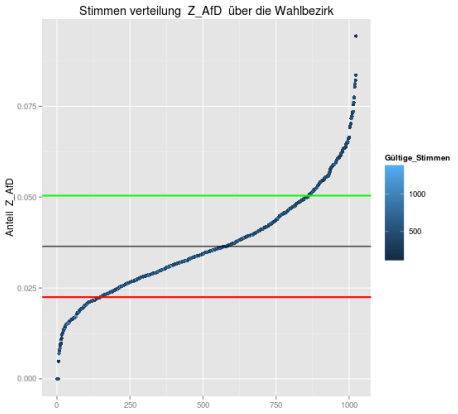
(c) Grüne



(d) Die Linke



(e) FDP



(f) AFD

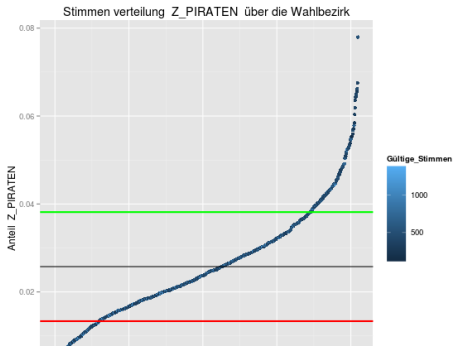
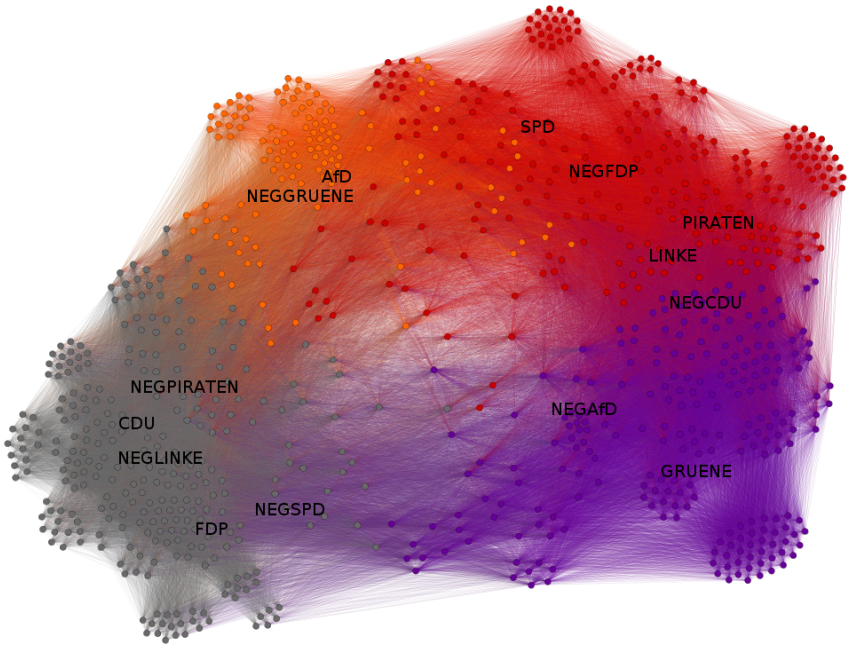
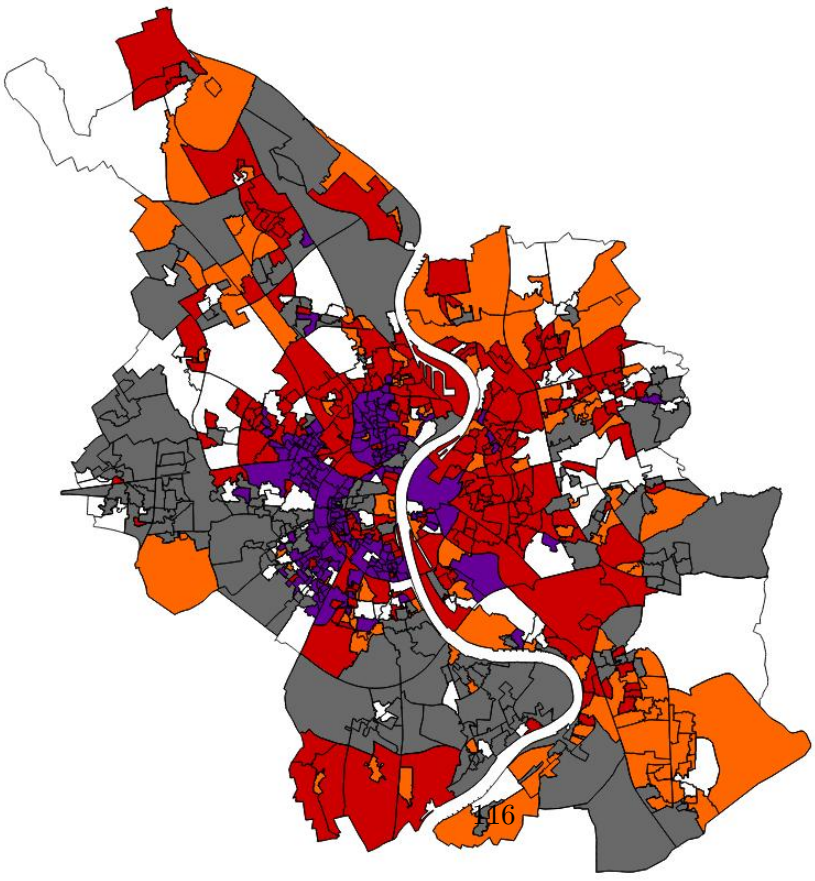


Abbildung 3.12: *Das Ergebnis der Wahldatenauswertung für die Stadt Köln. Die Clusterfarben sind in beiden Darstellungen gleich.*



(a) *Abbildung der Stimmbezirke als relationale Darstellung, durch ein Kräfte basiertes Verfahren angeordnet. Die Beschriftung zeigt die Parteipräferenzen an. Bezirke ohne spezifische Parteienpräferenz werden nicht abgebildet.*



(b) *Abbildung der Stimmbezirke als geografische Darstellung mit QGIS[205]. Die weißen Bezirke*

3.3.2 Ergebnis

Abbildung 3.12a zeigt eine Karte mit den Farben des Clusterings. Die Cluster sind nicht nur in der Netzdarstellung zu sehen, sie sind auch in der geografischen Darstellung sichtbar. In der Karte sind die "durchschnittlichen" Stimmbezirke, welche keine klaren Präferenzen aufweisen, weiß dargestellt.
Die Visualisierungen zeigen auch, dass kleinere, zentrale Stimmbezirke anders wählen als große Stimmbezirke am Rand der Stadt. Des Weiteren sind Stimmbezirke augenscheinlich oft ihren Nachbarn ähnlich.
Auffällig ist auch die Flächenverteilung in Abbildung 3.12a. Der lila Block ist sehr groß und die Fläche in Abbildung 3.12b eher klein, was an der geringeren Größe der Wahlbezirke bzw. der höheren Bevölkerungsdichte in den Innenstadtvierteln liegt.

3.3.3 Städtevergleich

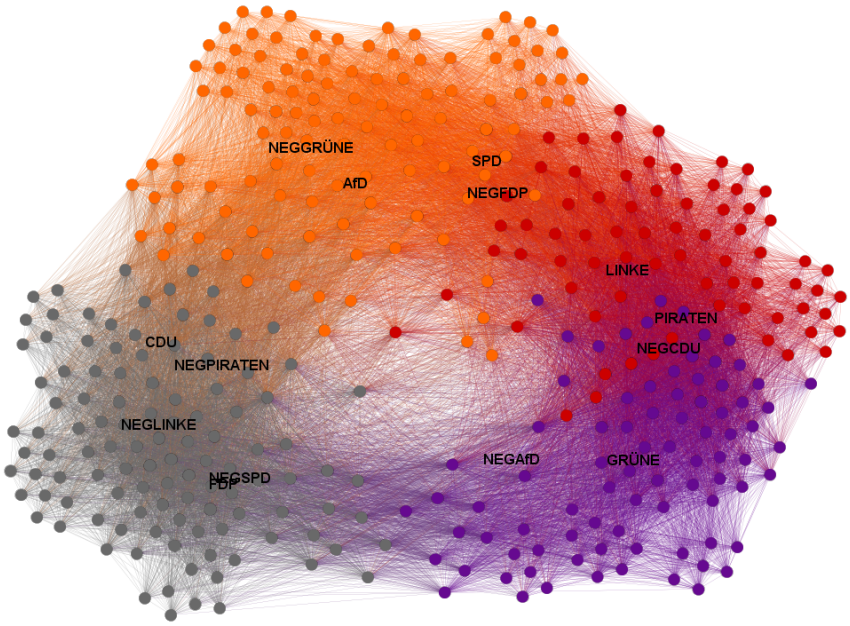
Das gleiche Vorgehen wie mit den Daten aus Köln lässt sich auf andere Städte mit gleicher Datenlage anwenden.
In diesem Fall hat Stuttgart die gleichen Wahldaten wie Köln bereitgestellt. Die Visualisierungen sind daher vergleichbar mit denen aus Köln.[19]
An der Abbildung ist zu sehen, dass der Clustering-Algorithmus ähnliche Gruppen fand, diese aber unterschiedlich groß sind. So ist hier die Gruppe der Orangenen auch mehr auf die SPD-Stimmen angeschlagen und das orangene Cluster wurde damit größer als in Köln. Die Verteilung der Wahlpräferenzen ist hier augenscheinlich anders strukturiert als in Köln.
Zum Vergleich dienen dabei die Abbildung 3.12b und 3.13b. In Köln ist eher eine Ost-West-Verteilung der Cluster zu sehen. Im Stadtkern ist das lila Cluster am speziellsten. In Stuttgart ist eher eine Nord-Süd-Verteilung der Cluster ersichtlich. Dies wird im Vergleich von Abbildung 3.13a und 3.12b ersichtlich.
Eine interessante Parallele zur nicht geografischen Visualisierung ist dabei das Loch in der Mitte der Visualisierung. Diese Position ist durch die Verteilung der Präferenzen zu erklären, widersprüchliche Präferenzen schließen sich aus. Zudem ist an dieser Stelle auch ein Loch in den Daten, da schwache Präferenzen ausgeschlossen wurden. Die Stimmbezirke, die hier sind, verbinden ansonsten eher widersprüchliche Wahlpräferenzen.

3.3.4 Der Ort als verbindende Größe

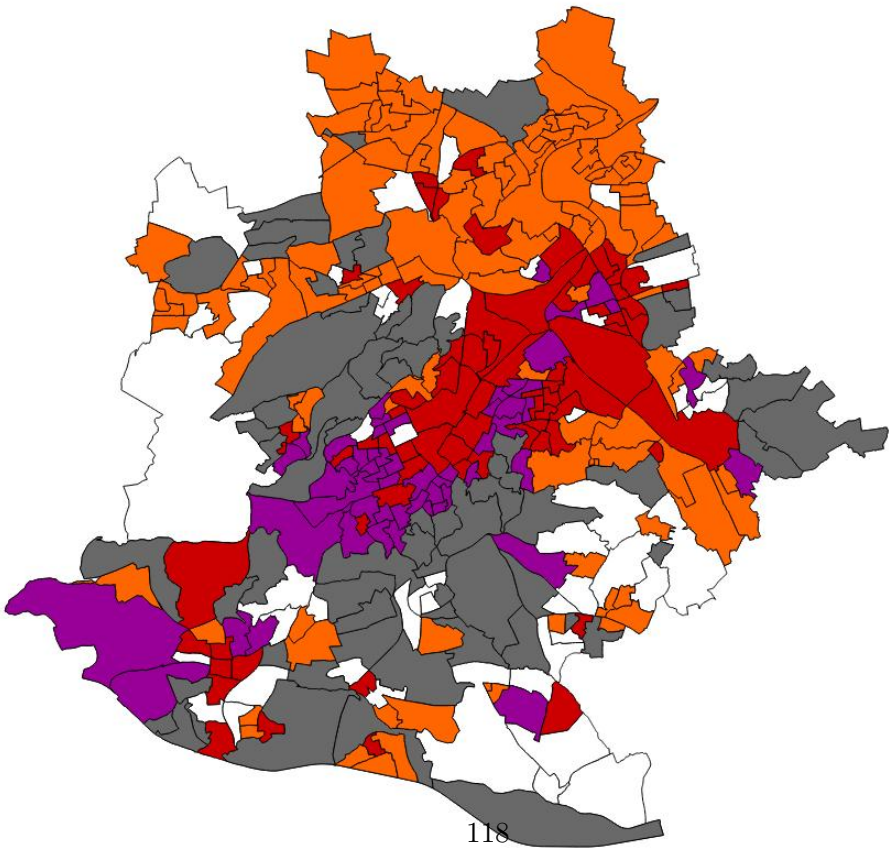
Die Wahlpräferenzen wurden bis jetzt als verbindende Größe zwischen den Stimmbezirken gesehen. Alternativ könn auch die Stimmbezirksrelation zwischen den Wahlpräferenzen betrachtet werden.
Im Fall von Stuttgart bildet sich eine Aufteilung in der Form, wie sie auch bei den Stimmbezirken in Abbildung 3.14 gezeigt wird. Der Clustering-Algorithmus bietet hier aber eine

[19] Da kräftebasierte Algorithmen die Beziehungen untereinander abbilden und nicht die gleichen Dinge an die gleichen Positionen setzen, musste die Ansicht gedreht werden. Dabei sind die Positionen und Abstände der Knoten untereinander, wie sie vom Algorithmus generiert werden, nicht verändert worden.

Abbildung 3.13: *Das Ergebnis der Wahldatenauswertung für die Stadt Stuttgart. Ansonsten wie Abbildung 3.12.*

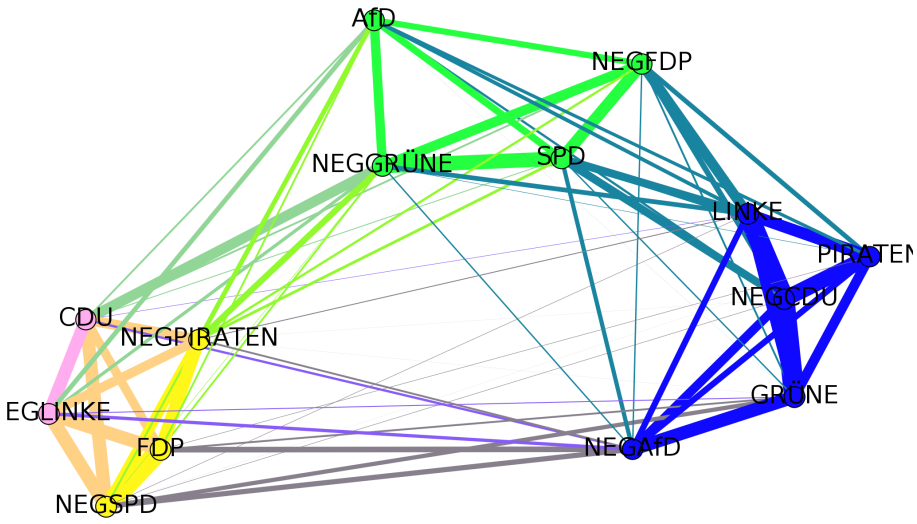


(a) *Abbildung der Stimmbezirke als relationale Darstellung für Stuttgart. Ansonsten wie Abbildung 3.12a.*



(b) *Abbildung der Stimmbezirke als geografische Darstellung für Stuttgart. Ansonsten wie Abbildung 3.12b.*

Abbildung 3.14: *Die Verbindung der Parteipräferenzen über die Stimmbezirke für die Stadt Stuttgart*



klarere Definition der Cluster. Die Präferenzen konnten beim Clustering der Stimmbezirke nur als etwaige Tendenzen gesehen werden. Auch hier wurde der Clustering-Algorithmus so parametrisiert, dass 4 Cluster enstehen. Im Gegensatz zum Ergebnis der Stimmbezirke können Präferenzen klar zugeordnet und gruppiert werden. Die Ergebnisse des Clusterings sind nicht vergleichbar, da eine andere Klasse von Knoten klassifiziert wird. Die Farben der Cluster wurden daher auch nicht angepasst, um die Verwirrung beim Vergleichen der beiden Darstellungen zu vermeiden. Die hier abgebildeten und klassifizierten Knoten bildeten die Kanten in der anderen Projektion.
Die Kompression zeigt den verbindenden Einfluss der Stimmbezirke, also den geografischen Zusammenhang zwischen den Wahlpräferenzen.
Der aus der Projektion erstellte Graph hat eine sehr hohe Dichte von über 80 Prozent. Das ist in diesem Fall aber auch die maximale Dichte des Graphen. Durch die Unmöglichkeit der Kanten zwischen negativen und positiven Items derselben Partei ist die maximale Anzahl der möglichen Kanten kleiner als in einem "normalen" Graphen.
Der Graph aus Köln dagegen weist eine andere Verteilung von Gewichten auf. Die Gewichte sind aufgrund der Menge der Daten stärker. In Köln gibt es etwa doppelt so viele Stimmbezirke wie in Stuttgart. Beim Akkumulieren ensteht daher eine höhere Gesamtsumme der Gewicht aller Kanten. Die Netze sind beide gleich dicht, daher verteilt sich das Gesamtgewicht auf die gleiche Anzahl von Kanten. Dies kann durch die Parametrisierung des Visualisierungsalgorithmus oder die Normalisierung der Kantengewichte abgefangen werden.

3.3.5 Fazit

Die hier entstandene Visualisierung setzt die Wahlpräferenz in zwei verschiedene Positionskontexte. Der erste Kontext ist die geografische Lage der Knoten. Der zweite Kontext ist die Position der Wahlpräferenz in Abhängigkeit der Wahlpräferenz aller anderen Bezirke.
Die Einfärbung dient als Brücke zwischen den beiden Darstellungen. So lassen sich zwei verschiedene Anordnungen im zweidimensionalen Raum miteinander vergleichen.
Auch ein Bezeichnungssystem auf Textebene wäre möglich. Dies wird jedoch, wie im Fall der Kölner Daten, bei 1024 Stimmbezirken sehr unübersichtlich. Mit der Überlagerung der Labels an den Knoten würde dies in einer starken Überlagerung der verschiedenen Bezeichner enden. Die Verwendung der relativen Positionierung der Parteipräferenzitems hat mit wenig Text einen Zugang zu den Clustern gegeben.
Die nicht geografische Darstellung kann in diesem Fall als eine Art Legende zum Leseverständnis der Kartendarstellung dienen.
Im Rahmen von dynamischen Visualisierungen könnten alternative explorative Methoden in die Betrachtung einbezogen werden. Die Verbindung der Knoten zwischen den Bildern ist explizit bekannt. Nur kann sie in einem rein statischen Kontext nicht interaktiv aufgelöst werden. So könnte durch das Anwählen der Knoten in der geografischen Visualisierung der Knoten in der abstrakten Visualisierung hervorgehoben werden und umgekehrt.

3.4 Datenbanken

Das Handhaben verschiedener Daten in einzelnen Dateien ist unhandlich und Fehleranfällig. Eine Menge von Informationen muss immer zusammengehalten werden. Dabei müssen Daten manuell gehandhabt und auch reproduzierbar bleiben. Ein Verbinden einzelner Datensätze ist auf Dateiebene schwer zu realisierren und zur Abfrage oder Analyse müss die Daten vollständig eingelesen werden.
Alle diese Probleme werden durch das Verwenden von Datenbanken gelöst. Datenbanken verwalten große Datenmengen und Datenbank-Management-Systeme schaffen einen gemeinsamen Zugriffspunkt auf Daten. Das ist wichtig, wenn viele verschiedene Nutzer auf eine Datenbasis zugreifen sollen. Es ist ein Mechanismus, um bei gleichzeitigen Änderungen keine Inkonsistenzen in den Daten entstehen zu lassen. Es wird auch sichergestellt, dass Daten einem Schema folgen und dass Daten überhaupt in die Struktur von Datenbanken passen.

3.4.1 Datenbanktypen

Eine Datenbank an sich ist ein Programm, das das Ablegen und den Zugriff auf Daten ermöglicht. Gegeben werden auch einfache strukturierende Elemente oder eine Datenhaltungsphilosophie. Diese Grundstrukturen können für die Modellierung eines eigenen Datenmodells genutzt werden. Datenbankensysteme folgen meistens einem spezifischen Paradigma der Datenstrukturierung.

Relationale Datenbanken

Relationale Datenbanken gibt es seit den 80er Jahren. Es gibt daher potenziell eine Vielzahl an bestehenden Datenbanken, die seitdem angelegt wurden. Dabei sind die Strukturen in relationalen Datenbanken vom Prinzip auf Relationen angelegt, auch wenn das Relationsmodell nicht auf Netze und Verbindungen angepasst ist.
Ein Datensatz kann im relationalen Schema als eine referenzierbare Menge von Werten beschrieben werden. Datensätze des gleichen Typs können die gleichen Werte zugewiesen werden. Datentypen werden in Tabellen angelegt. In den Spalten dieser Tabellen stehen vergleichbare Werte und in den Zeilen befinden sich die bennbaren Datensätze. Datensätze werden durch den Primärschlüssel und den Tabellennamen eindeutig einzigartig benennbar. Die Darstellung als Tabelle hat den Vorteil, dass Werte derselben Eigenschaft untereinander stehen.
Bis hierhin hat eine Datenbank aus Datensatztupeln wenige Vorteile gegenüber einer Tabelle, wie sie auch auf Klemmbrettern und in Aktenschränken Platz finden würde. Doch sind relationale Datenbanken auch in der Lage, Daten in Relation zueinander zu setzen und diese Relationen auf zu lösen. Die Relationen werden durch Fremdschlüssel und Referenz- oder Kreuztabellen[20] hergestellt. Welche Methode verwendet wird, hängt davon ab, welche

[20] Engl.: reference table

Kardinalität eine Beziehung hat. Diese Kardinalitäten können in drei Gruppen unterteilt werden.

1 zu 1 Genau ein Datensatz von Typs A kann einem anderen Datensatz von Typs B zugeordnet werden. Ein Fremdschlüssel des anderen Datensatztyps wird dafür hinterlegt.

1 zu N Ein Datensatz von Typ A kann einer Menge anderer Datensätze von Typ B zugeordnet werden. Hier wird für gewöhnlich ein Fremdschlüssel von Typ A in der Tabelle von Typ B hinterlegt.

N zu M Ein Datensatz kann vielen anderen Datensätzen zugeordnet werden und umgekehrt. Eine Referenztabelle enthält einen einen Fremdschlüssel auf Typ A und einen auf Typ B.

Die Kardinalität einer Verbindung gibt an, wieviele Datensätze des einen Typs mit einem anderen Typen verbunden werden können. Die Kardinalität legt dabei nicht fest, welche Bedeutung eine Beziehung hat.[211]

Graphdatenbanken und Triple Stores

"Relational Databases Lack Relationships"[212]
Eine Graph- oder Netzwerkdatenbank ist eine Datenbank, die im Design dem Netzwerkmodell folgt. Sie ist im Gegensatz zum relationalen Datenbankmodell auf Beziehungen spezialisiert.
Da das eigentlich relationale Datenbankmodell doch sehr unintuitiv ist wenn es um Relationen geht, gibt es einige Datenbankmodelle, die direkt auf ein Netzmodell aufsetzen, beispielsweise Neo4J. Neo4J ist direkt auf die Abbildung von Graphen-Strukturen ausgelegt.
Desweiteren gibt es Triple Stores. Triple Stores sind Datenbanken für RDF-Daten. RDF ist ein komplett auf typisierten Relationen ausgelegtes Modell, das anders als Neo4J, Knoten nicht mit Datensätzen oder Dokumenten, sondern nur mit weiteren Relationen auf Datenfragmente beschreibt. Durch diese sehr granulare Datenstruktur müssen viele Subjekt-Prädikat-Objekt Aussagen verwaltet werden.
Neben der internen Struktur ist auch die Abfrageschnittstelle ein zentrales Element von Datenbanken. In relationalen Datenbanken die über SQL abgefragt werden, können Relationen über "JOIN"-Befehle ausgewertet werden. Relationen können über diese Join Konstrukte nur aufwendig abgefragt werden. In anderen Sprachen, wie Cypher für die Datenbank NEO4J und SPARQL für RDF-Datenbanken werden Relationen durch weniger syntaktischen Aufwand abgebildet.
In relationalen Datenbanken ist es auch sehr Umständlich neue Beziehungstypen hinzuzufügen. Dazu muss die Struktur der Datenbank angepasst werden. In Graphdatenbanken ist das nicht der Fall.[21]

Dokumentdatenbanken

Dokumentdatenbanken beziehen sich auf Datenbanken, die einzelne Datensätze mit individueller Struktur erlauben. Datensätze haben anders als bei der relationalen Datenbank

keine starre Struktur. Eigenschaften können anders als in relationalen Datenbanken mehrere Werte beinhalten.

XML-Datenbanken und Content Management Systeme

XML-Datenbanken bieten einen Mantel aus Funktionen um XML wie XQuery und XPath. XML an sich ist eigentlich schon eine komplexe Datenstruktur, in dessen Umfeld standardisierte Transformations- und Validierungsstandards existieren. XML-Datenbanken sind daher eher ein Verwaltungswerkzeug. Meist ist das Aufkommen von annotiertem Text sehr überschaubar und kann von den meisten Rechnern ohne Probleme im Arbeitsspeicher ablaufen. Datenbanken wie eXist[166] stellen dabei Möglichkeiten zum Abfragen mit XPath oder XQuery bereit. Die Abfrage mit XPath-Ausdrücken sind aber nach einmaligem Parsen einer Datei mit einem Programm oder einer Programmbibliothek, wie beispielsweise BaseX[77, 113] oder Xeletor[270], möglich.
HTML- und XML-Daten werden auch schnipselweise in normalen Textfeldern in relationalen Datenbanken gehalten. Das ist oft im Kontext von Content Management Systemen, wie Drupal[78] der Fall.

3.4.2 Datenmodell

Die Modellierung von Datenbanken ist ein Prozess, der versucht, einen Sachverhalt abzubilden. Dies hat zur Folge, dass eine Datenbank nur bestimmte Informationen in Relationen umsetzen kann. Dies könnte als Meta-Netz bezeichnet werden. Dieses gibt dabei vor, welche Relationen zwischen welchen Dingen geschlossen werden können. Mit dem Datenmodell einer Datenbank, werden die geistigen Weichen gestellt, in denen ein Sachverhalt abgebildet werden kann.
Ein Datenbank-Modell spiegelt also auch immer eine Sicht auf ein Thema oder einen Sachverhalt wider. Manche Dinge können in Relation gesetzt werden und andere nicht.
Werte und Merkmalsausprägungen stellen eine Selektion dar was überhaupt erfassbar ist. Was nicht abgebildet wird, ist damit für diese Datenbank nicht von Belang. Um den abgebildeten Sachverhalt bildet sich eine geschlossene Welt. Alle Anfragen auf Daten, die in der Datenbank nicht verfügbar sind, werden mit “Falsch” beantwortet, auch wenn die Antwort auf die Anfrage genaugenommen unbekannt ist.
Die Struktur einer wirklichen Datenbank kann im Gegensatz zur konzeptionellen Datenbank, repräsentiert in einem Entity-Relationship-Modell (ERM), eine andere Form aufweisen. Eine abgebildete Relation muss bei entsprechender Kardinalität von N:M auch durch eine Tabelle abgebildet werden. Für einen Relationstypen ist es hier von Interesse, ob diese Relation auch eine Intention oder andere Werte enthält. So kann eine Referenztabelle auch mit weiteren qualifizierenden Informationen bestückt sein.
Das entspricht den Gewichten oder Labels auf einer Graphenkante in einem multipartiten Graphen. Aus einer solchen Struktur kann man ein Meta-Netz erstellen, welches die Möglichkeiten der Modellierung widerspiegelt.

Identität

Generell haben alle Datensätze in allen Datenbanktypen einen eineindeutigen Bezeichner, einen sogenannten Schlüssel. In Datenbanken wird ein solcher Schlüssel oft durch eine ganzzahlige Nummer abgebildet. Das hat den Vorteil, dass Computer mit Zahlen sehr schnell Vergleiche anstellen können und ein minimaler Speicher bei der Abbildung verbraucht wird.[21] In einer relationalen Datenbank darf ein Schlüssel jeweils nur einmal in einer Tabelle vorkommen. Aber wenigstens für die Datenbank-Instanz ist der Schlüssel eindeutig. Daraus folgt, dass immer der Kontext gebraucht wird, um mit einem Schlüssel etwas identifizieren zu können.

Strukturelle Identität

In der relationalen Datenbank-Theorie können bestimmte Normalformen angesetzt werden. Die Normalisierung gibt an, wie stark eine Datenbank zum relationalen Geflecht wird, um Redundanz zu vermeiden. Bei einem geringeren Normalisierungsgrad werden Daten eher in einer Tabelle abgebildet.
Je weiter also die Normalisierung einer Datenbank fortschreitet, umso mehr werden Datensätze aufeinander verweisen. Die Struktur wird komplexer und um ein Ansichtsergebnis zu erzeugen, werden mehr Einzeldatensätze zusammengefügt. [211, Kapitel 5] Eine Entität kann also in mehreren Tabellen und Datensätzen beschrieben werden, wenn es aus mehreren Teilen besteht. Die Normalisierung ist eigentlich ein Konzept für relationale Datenbanken, doch in anderen Datenbankformaten lassen sich ähnliche Überlegungen zur Granularität vornehmen.
Die Identität und der Zusammenhang von Datensätzen lassen sich anhand von Constraints erschließen. Beim Löschen oder Ändern von Datensätzen werden mit Constraints auch anhängende Datensätze geändert oder gelöscht. Sie geben implizit Aufschluss darüber, welche Datensätze nur im Kontext von anderen Datensätzen Sinn ergeben und welche Datensatztypen für sich selbst interessant sind.

Tags, Vokabulare

Freitext kann in einer Datenbank ein Problem darstellen. Jeder Eingebende würde eigene Worte verwenden. Dieses Problem kann durch ein kontrolliertes Vokabular gelöst werden. Ein Datenbankfeld wird dabei auf Begriffe aus einer Liste beschränkt. Das vermeidet dabei Fehler wie Rechtschreibfehler oder Homonyme, die bei der automatischen Auswertung die Vergleichbarkeit von Datensätzen stört. Eine solche Festschreibung von Begriffen auf ein Datenbankfeld nennt man kontrollierte Vokabulare. Ähnlich funktioniert auch Tagging. Beim Tagging können einzelne Datensätze Tags zugeordnet werden. Diese Tags sind statisch und dienen der Beschreibung.
Der größte Unterschied liegt darin, dass ein Tag von einem Nutzer “bottom up” vergeben

[21] Je nach Datenbank kann sich ein Schlüssel auch aus der Kombination mehrerer Felder zusammensetzen. Wichtig ist, dass ein Primärschlüssel einzigartig ist.

wird. Ein kontrolliertes Vokabular wird “top down” vorgegeben und lässt dem Nutzer keinen Freiraum für eigene Vorschläge. Diese Zuordnungen können auch in bipartite Netze umgewandelt werden und dann weiter in Ähnlichkeitsnetze projiziert werden.

3.4.3 Web und Datenbank

Hinter fast jeder Website oder Webanwendung steht eine Datenbank. Mit dem Paradigma des Web 2.0 ist das WWW dynamischer geworden. Websites[22] bestehen in ihrer klassisch statischen Form nicht mehr so zahlreich.
HTML wird immer noch als technischer Standard verwendet, doch bieten diese menschenlesbaren Repräsentationen nur noch die Hülle für Informationen. Die Inhalte werden oft dynamisch oder teildynamisch beim Aufruf konstruiert. Dabei stehen Datenbanken im Hintergrund und dienen als Rohmaterial für die Darstellung und auch für die Verlinkung von Websites. Des Weiteren kann eine Website dynamisch, je nach aufrufendem Computer, andere Inhalte ausliefern. Das stellt ein Problem für Hypertextbewertungssysteme wie beispielsweise Page-Rank dar.[23]
Auch bei Content Management Systemens und Wikis steht fast immer eine Datenbank im Hintergrund.[24] Das hat unter anderem den Grund, dass Datenbanken meist besser zu pflegen sind als statische HTML-Dateien. Dazu kommt, dass Datenbanken die Bearbeitung mit mehreren Personen meist nativ unterstützen. Änderungen können von mehreren Personen parallel und konsistent vorgenommen werden. Zudem können beliebige Metadaten und spezielle Verwaltungsfunktionen abgewickelt werden, wie beispielsweise die Verwaltung von Übersetzungen von Texten.
Im Falle eines Content Management System ist eine Datenbank der Informationsträger für die Informationen, die auf einer Website präsentiert werden sollen. Die Datenbank ist hier Mittel zum Zweck. Ihre Struktur zielt darauf ab, den Inhalt von HTML-Seiten besonders gut zu verwalten. Die Planung und Modellierung ist damit generisch. Der Sinn ist es, Daten im Web darzustellen und diese zu verwalten.
Eine Datenbank kann aber auch eine intentionale Zusammenstellung oder Aggregation von Wissen sein. Das Web-Frontend muss dann die Funktion erfüllen, dass diese gesammelten Daten nach außen menschenlesbar dargestellt werden. Die Struktur der Datenbank wird dann durch einen anderen Zweck bestimmt als die reine Repräsentation von Inhalten.
Allgemein ist aber ein Web-Frontend immer ein intentionaler Filter. Die vorhandenen Informationen werden kombiniert und je nach Anfrage kann sich der Inhalt einer Seite anders darstellen. Beispielsweise könnte ein zufälliger Text eingeblendet werden. Andererseits können individuelle Angebote angezeigt werden, wie bei personalisierter Werbung. Je nach Nutzer, Herkunft, Uhrzeit, Zielgerät können Inhalte anders oder überhaupt nicht gezeigt werden.

---

[22] Die Abgrenzung zu Hypertext und reinem HTML ist hier sehr bewusst gewählt, da Hypertext/Hypermedia und HTML immer etwas Statisches implizieren.

[23] Die Verarbeitungsweise von Page-Rank wurde in Abschnitt 2.2.3 beschrieben.

[24] Es gibt auch sehr simple Systeme für den einfachen Zugriff auf die Daten wie z.B. Tiddly wiki. In diesem System werden die Daten in Dateien gespeichert.

3.4.4 Von der Datenbank zum Netz

Aus einer Nutzerperspektive kann in einer Datenbank meist ein einzelner Datensatz mit seinen Kontexten betrachtet werden, beispielsweise ein Artikel in einem CMS, eine Wiki-Seite, Informationen zu einem Menschen oder Informationen einem Objekt. Dazu werden Relationen oder Informationspartikel dargestellt wie die Klassifikation durch ein Tagsystem, Eigenschaften, die mehr als einmal auftreten können, wie die Zuordnung von Links, Leserkommentaren u.ä. Das entspricht meistens einem Datenbankreport.
Diese Darstellungsweisen sind mit einem egozentrierten Netzwerk vergleichbar.
Standardsuchen nach Suchbegriffen oder facettierte Suchen können mehrere Datensätze gleichzeitig anzeigen. Das Resultat ist meist eine gewichtete Liste der Kurzinformationen zu Datensätzen. Die Gewichtung schlägt sich in der Reihenfolge nieder. Bei keiner dieser standardisierten Vorgehensweisen wird eine Netzstruktur aufgezeigt, obwohl Netze in den Daten enthalten sein können.
Bei der Umwandlung von Datenbanken zu analysierbaren Netzen können verschiedene Wege gegangen werden. Dabei ist es auch wichtig, welche Eigenschaften das Potenzial haben Netze zu bilden und welche Verbindungstypen nur Zuordnungen beinhalten. Auch muss abgeschätzt werden, wie viel Aufwand und menschlicher Eingriff nötig ist, um Netze in Daten zu erkennen.

Relationen

Die einfachste Möglichkeit eine Datenbank in ein Netz um zu wandeln ist, alle Datensätze als Knoten und alle Relationen als Kanten abzubilden. Dieses Netz wäre von den Knotentypen her divers. Der gemeinsame Nenner der Knoten wäre, dass sie alle “Datensätze” repräsentieren.
Je nach der Intention einer Datenbank bieten sich Möglichkeiten zur Analyse des Gesamtzusammenhangs der Datenmenge. Eine Datenmenge, die bei Betrachtung aller Relationen eine komplette oder sehr große Komponente bildet, hat einen sehr viel integrativeren Charakter als eine Datenbank, deren Daten kleine Komponenten bilden. Eine Analyse der Komponenten in einer Datenbank kann dabei verwaiste Teile des Datenbestandes zeigen, beispielsweise als einzelne unverbundene Komponenten.
Die Betweeness Centrality könnte bei diesem generischen Ansatz die Wichtigkeit eines Datensatzes für die Gesamtstruktur der Datenbank liefern. Interessant für Datenbanken im Allgemeinen kann eine solche Zentralitätsanalyse werden, um herauszufinden, welches die verbindenden Stücke einer Datenbank sind. Dabei können hier strukturelle Probleme aufgezeigt werden. Wenn eine Art von Datensatz viele verschiedene Arten von Relationen mit hohen Kardinalitäten zulässt, dann ist potenziell eine starke Vernetzung von diesen Datentypen gegeben. Das muss aber in der Praxis nicht der Fall sein. Es kann eine Diskrepanz zwischen der aus dem Modell hervorgehenden und der realen Struktur geben. Das heißt, dass Teile einer Datenbank in der Praxis sehr selten oder sogar überhaupt nicht genutzt werden.
Bei einer Betrachtung der unterliegenden Daten muss beachtet werden, dass nicht alle Datensätze auch für sich alleine einen interessanten Fakt darstellen. Sie können je nach

Normalisierung auch Informationsfragmente bilden, die für sich genommen unverständlich sind. Das kann bei sehr granularen Daten, kontrollieren Vokabularen und Tags der Fall sein. Ein kleiner Informationsschnipsel, der oft verwendet wird, kann eine zentrale Rolle vortäuschen, obwohl er eigentlich nur generisch ist.

Relationstypen in relationalen Datenbanken

Datensätze sind in relationalen Datenbanken durch ihre Tabellenart zwangsweise typisiert. Eine Tabelle muss immer von anderen differenzierbar sein. Dabei sind die Datensätze aus der gleichen Tabelle auf Eigenschafts-Ebene anhand der Wertausprägungen vergleichbar. In klassisch relationalen Datenbanken wie MySQL sind das die Datenbank Felder des gleichen Typs.
Die Relation ist dabei nicht zwangsweise typisiert. Der eigenschaftslose Link kann auch mithilfe der Typen der Endpunkte typisiert werden, wenn der Typ des Links nicht dokumentiert oder klar definiert ist.
Eine klare Strukturierung von Daten und Sachverhalten kann auch von der Nutzung und der Aufgabe einer Datenbank abhängen sowie von der Konsistenz ihrer Nutzung. Eine Datenbank, die klar geplant wurde, wird konsistenter sein als eine Datenbank, die erweitert wurde oder nicht nur einem klar definierten Zweck dient.
Zeigen lässt sich dieses Problem an einem klassischen Beispiel aus Lehrbüchern der Datenmodellierung, der Verwaltung von Studierenden und Seminaren. Solange die Datenbank nur diesem Zweck dient und nichts anderem, gibt es die Relation zwischen Student und Veranstaltung. Der Student besucht das Seminar.
In bestimmten Fällen kann aber gerade bei einer Veränderung und Erweiterung des Funktionsumfangs einer Datenbank eine Überlagerung verschiedener Bedeutungen erzeugt werden, falls diese nicht explizit modelliert wurde. Das ist im Besonderen der Fall, wenn eine Verbindung nicht weiter definiert ist.

Nietzsche in der Student-Seminar Datenbank

Zur Verdeutlichung der Klassifikationsprobleme von Datenbankrelationen wird das Lehrbuch Beispiel zugrunde gelegt in dem Studenten einem Seminar zugeordnet werden. Nehmen wir an im 19. Jahrhundert hätte es an der Universität Bonn einen Zettelkasten, als Vorform einer Datenbank gegeben. Zu diesem Zeitpunkt wird ein junger Student namens Friedrich Nietzsche, wie viele andere Studenten, einem Seminar der Philosophie zugeordnet.
Im Jahr 2000 wird der Zettelkasten digitalisiert und zur Datenbank. Im gleichen Atemzug wird erkannt, dass Studierende auch Menschen sind, und abstrahiert alle Studenten in der Datenbank zu “menschliche Wesen”. Ein Vermerk im Menschendatensatz markiert sie als Studierende, DozentIn, ProfessorIn oder großeR DenkerIn. Die Datenbank wird um das Thema von Vorlesungen erweitert. Menschen die Philosophen sind, wie Aristoteles und Plato, werden Seminaren zugeordnet. Dies geschieht, um die Seminare thematisch einzuordnen.

Dabei kommt es auch wieder dazu, dass Nietzsche einem Seminar zugeordnet wird.
Die Verbindung von Nietzsche zum Seminar ist jetzt mindestens zweideutig. Nietzsche ist Mensch, Student und Philosoph und daher Besucher und Thema zugleich. Er war als Student Teilnehmer eines Seminars und in einem anderen Thema des Seminars.
Die Verwirrung lässt sich mit einfachem Menschenverstand lösen. Doch bleibt es ein Problem bei der sauberen Abbildung von Daten. Die Verbindungen müssten typisiert werden und nicht die Menschen. Die Rolle des Studierenden ergibt sich aus seiner Rolle in Beziehung zum Seminar.

Freitext

Auch in Datenbanken gibt es Text. Dies kann von kontrollierten Vokabularen, über kurze Beschreibungen zu ganzen Buchbänden reichen. Die Menge und der Umfang von Text sind dabei abhängig von der Art der Datenbank. Datenbanken auf Grundlage von Content Management Systemen werden mehr Text enthalten. Die Menge des Freitextes ist dabei auch von der Modellierung und dem Thema abhängig. Die Menge von Datenbank-Texten wird dabei voraussichtlich eher granularer sein als die Menge an Text, die sich in reinen Hypertextdokumenten versteckt.
Wichtig ist, was die Knoten und die Kanten bilden soll. Hier ist wieder die Fragestellung entscheidend.
Ein einfaches Maß einer Übereinstimmung oder Ähnlichkeit eines Dokumentes kann durch die Ähnlichkeit der verwendeten Worte genommen werden. Hier wären beispielsweise Maße wie TF-IDF[25] zu nennen.

Umwandlung, Homomorphismus und Vorteile von Datenbanken

Die Abschnitte über Datenbanken, statische Daten, annotierte Texte und Textverwaltungssysteme lassen sich nicht direkt als unvereinbar verstehen. Die Formate lassen sich teilweise ineinander überführen. So lassen sich Wikis wie die Mediawiki in statische XML-Dateien verwandeln. Diese sind vollständig in der Dokumentation, aber sie funktionieren nicht. Auch lassen sich CMS-Systeme oft vollständig in eine statische HTML-Version übersetzen. Tabellen können in Datenbanken überführt werden und umgekehrt. Die Untersuchung als Netz stellt eine spezielle Auswahl und Akkumulation dieser Daten dar.
Datenbanken und deren Schnittstelle zum Menschen beruhen auf einer sinnvollen, intentionalen Auswahl und Zusammenstellung. Interessante Dinge sollen sichtbar werden und uninteressante Dinge wegfallen oder als Einzelfälle aufkumuliert werden.

3.5 Patterns im automatischen Retrieval

In diesem Abschnitt wird ein Verfahren beschrieben, das hilft, Netze in Relationssystemen von Datenbanken zu erkennen. Im Mittelpunkt steht dabei die Matrixprojektion von bipartiten Netzen in Referenztabellen von Datenbanken.

[25] Eine Beschreibung findet sich in Abschnitt 3.2.1.

Es wird nicht auf das intentionale Einzelergebnis aus den intentionalen Ansichten geschaut, sondern versucht, automatisch alternative Ansichten auf Daten zu finden, die ansonsten nicht auf diese intentionale Art zusammengesetzt werden. Die Auswahl, ob es sich um sinnvolle Zusammenhänge handelt, wird anhand der graphentheoretischen Maßzahlen und Visualisierungen bestimmt. Hier wird versucht, ein exploratives Vorgehen mit wenig Vorwissen zu generieren, das sich an der tatsächlichen Struktur der Daten innerhalb seiner Modellierung orientiert und weniger am Modell selbst.
Das Vorgehen hat Ähnlichkeiten mit dem Meta-Matrix-Verfahren nach Diesner und Carley.[72, 74] Der Meta-Matrix-Ansatz[26] versucht Entitäten in Texten zu erkennen und diese über das gemeinsame Auftreten im Text in Beziehung zu setzen. Dabei entstehen bipartite Netze aus den Kombinationen der gefundenen Entitätstypen. Die Möglichkeiten der Kombinationen ergeben sich aus den extrahierten Entitätstypen. Diese Netze werden dann auf unipartite Netze projiziert und ihre strukturellen Eigenschaften werden auf Auffälligkeiten untersucht. Die Projektionen haben den Vorteil, dass die Netze kleiner werden und nicht transitive Knoten wegfallen.
Die Abwandlung besteht darin, dass die Entitäten nicht extrahiert werden müssen, sondern schon in der Datenbank zu Analyse bereit liegen.
Dabei wird der Frage nachgegangen, wie dieses Verfahren genutzt werden kann, um folgende Aufgaben zu unterstützen:

1. interessante Strukturen und Fakten auf höherer Ebene zu finden, die bei der Eingabe nicht intendiert waren (Serendipität),
2. Gesamtzusammenhänge in heterogenen, zusammengetragenen Datenbeständen zu finden,
3. Strukturen in Datenbanken zu erläutern,
4. fehlerhafte oder problematische Datenstrukturen zu erkennen,
5. fehlerhafte oder problematische Datenmodellierung zu erkennen.

3.5.1 Erläuterung des Cooccurrence Verfahrens

In diesem automatischen Verfahren werden aus bipartiten Netzen unipartite Netze erzeugt. Diese Netze bestehen aus den Kombinationen von Tabellen in der Datenbank, die verschiedene Knotentypen charakterisieren.
Die Datenbankkategorien werden über ihre Verbindungen zueinander analysiert.
Es können bei diesem Verfahren auch Verbindungen gewählt werden, die einer bestimmten Semantik entsprechen. Wenn in einer Datenbank keine Semantik der Beziehungen hinterlegt ist, dann kann diese aus der inhärenten Kombination der verbundenen Datensatztypen hervorgehen. Das heißt, wenn Datensätze aus zwei Kategorien miteinander verbunden werden können, dann geht über den Kontext der archäologischen Datenbank eine implizite

[26] Im Original “meta-matrix approach”.

Semantik der Verbindung hervor.
Semantik ist für das Verfahren an sich nicht wichtig. Sie tritt dafür umso mehr bei der Interpretation des Ergebnisses durch den Menschen in den Vordergrund.

Leistung

Um die Komplexität von teilweise sehr großen bipartiten Netzen zu reduzieren, werden sie projiziert. Bei den Ausgangsdaten handelt es sich um Listen, wie sie für gewöhnlich aus Datenbankabfragen hervorgehen. Zur traditionellen Projektion müssen diese Listen in Matrizen verwandelt werden. Die bipartite Matrize hätte einen Speicherbedarf, welcher der Anzahl der Knoten auf Seite 1 multipliziert mit Anzahl der Knoten auf Seite 2 entspricht.
Die beiden projizierten Netze hätten eine Größe, die der Menge von Knoten auf der Projektionsseite zum Quadrat entspricht. Dies ist ein sehr speicherintensives Vorgehen, daher wird hier nativ auf den Listen operiert. Diese können direkt aus einer Datenbank exportiert werden. Dies hat den Vorteil, dass die Daten direkt verwendet werden können. So ist es möglich, große Mengen von Daten zu verarbeiten.
Im Gegensatz zur Matrixoperation, die die Transformation der Listen in Matrizen nötig machen würde, kann hier nativ auf den Listen operiert werden, welche einen sehr viel kleineren Speicherbedarf haben.
Die Ausgabe von Listen wird von den meisten Abfragesprachen nativ bereitgestellt. So hat MySQL, wie es in der Arachne Datenbank verwendet wird, einen einfachen Export von Abfrageergebnissen in Listen als CSV-Datei.

Algorithmus

Die folgende Beschreibung des Algorithmus bezieht sich auf die Abbildung 3.15. Die Projektion erfolgt über die runden Buchstabenknoten. Die Kanten in dieser Abbildung sind nach diesen Knoten eingefärbt. Die achteckigen Knoten werden im Ergebnis als Knoten überbleiben.
Als Eingangsformat des Algorithmus wird eine Liste von Schlüsseln oder Identifiern aus einer Referenztabelle genommen. Die Liste hat in der ersten Spalte die Knoten der ersten Gruppe des bipartiten Netzes. In der zweiten Spalte stehen die Identifier der zweiten Gruppe des bipartiten Netzes.
Die Eingangsdaten werden zu Anfang nach den Knotenbuchstabenknoten sortiert (1).
In der Praxis kann eine solche Sortierung auch direkt beim Abfragen der Datenbank vorgenommen werden. Im ersten Schritt wird diese Ausgangsliste durchlaufen. Alle Zahlenknoten, die mit Knoten A verbunden sind, kommen in ein Set(2). Ein Set ist voll, wenn alle Einträge durchlaufen sind, die eine Verbindung mit Knoten A darstellen. Da die Liste sortiert ist, braucht es nur einen Durchlauf der Liste, um alle Sets zu bilden.
Im nächsten Schritt werden alle Knoten in Set A miteinander Verbunden. Die Kanten, die so entstehen, werden in eine neue Liste geschrieben.
Nicht alle Knoten in der Projektion bilden Kanten. Knoten E beispielsweise ist nur mit einem einzigen anderen Knoten verbunden, er erzeugt daher keine Kanten im projizierten

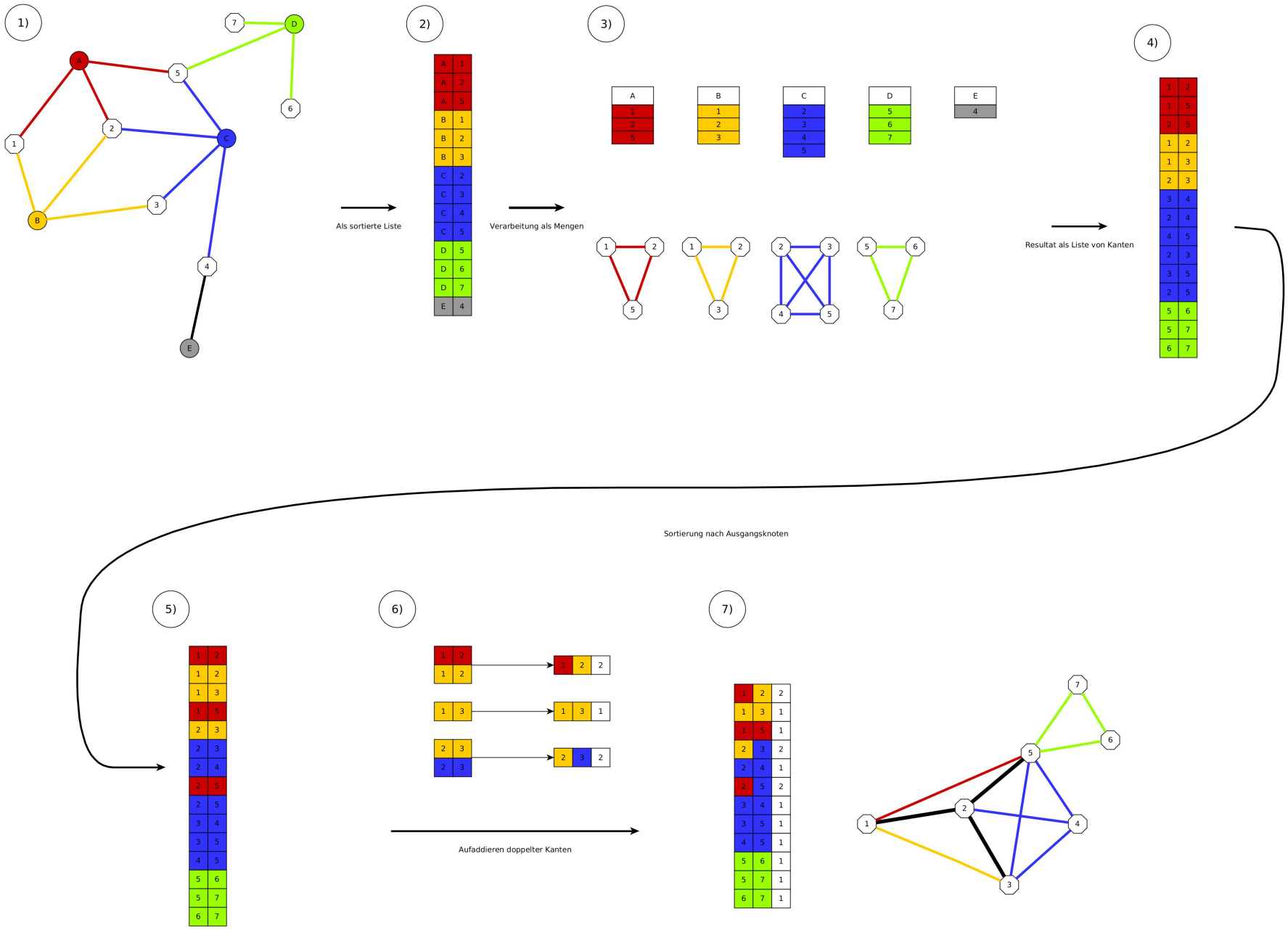


Abbildung 3.15: *Ablauf des Algorithmus zum Projizieren des eines bipartiten Graphen*

Graphen. Dies ist vergleichbar mit “prunning” bei Entscheidungsbäumen.
Im zweiten Schritt werden die neu erstellten Kanten in eine Liste geschrieben(4) und sortiert(5). Durch die sortiertirung der Liste können die parallelen Kanten in einem Durchlauf identifiziert werden. Die parallelen Kanten können dabei zusammengefasst werden(6). Bei der Zusammenfassung werden die neu erstellten Kanten gewichtet. In Punkt(7) sind die gewichteten Kanten als Liste und Graph aufgeführt. Die Farben in der Abbildung zeigen die ursprünglichen Knoten, aus denen die Kanten hervorgegangen sind. Kanten, deren Ursprünge sich überlappen, werden schwarze dargestellt.

Gewichte

Neben dem generellen Gewicht der parallelen Kanten, wie es in Punkt(7) in der dritten Spalte gezeigt wird, kann auch das hyperbolische Gewicht[181] der Kanten errechnet werden, im Weiteren wird der Wert bei der Ausgabe auch als Uniquity bezeichnet, da durch dieses Gewicht die Einzigartigkeit einer Verbindung abgebildet wird. Die berechnungsgrundlage für diese bewertung wird in Schritt (3) des hier vorgestellten Algorithmus gelegt.
Im Weiteren wird auch für jede Kante eine durchschnittliche Uniquity errechnet, dies geschieht in Schritt (6) und (7). Diese Maßzahl zeigt an, inwieweit eine Verbindung auf einzigartigen Verbindungen beruht. Sie lässt sich in einem Wert zwischen 1 und 0 ausdrücken, ihr Vorteil besteht darin, dass der Wert nicht mit steigendem Wissen ansteigt, sondern dass er immer zwischen 0 und 1 liegt. Dadurch wird verhindert, dass extreme abweichende Werte entstehen, wie sie beispielsweise bei besser erfassten Sachverhalten, mit Bias, in Datenbanken auftauchen.

Ausgabe

Das Programm gibt drei Informationsblöcke pro Netzwerk aus, den ersten für das bipartite Netz, das aus der Datenbank extrahiert wurde, die anderen beiden sind die Projektionen dieses bipartiten Netzes.
Die Zeilen der einzelnen Reports sind wie folgt strukturiert: Die erste Zeile zeigt die Art der Verbindungen an.
Dies sind im bipartiten Fall die verbundenen Datentypen. Beispielsweise für das Netz aus Literaturen und Objekten:

literatur <->objekt[27]

Im unipartiten Fall werden die projizierten Komponenten angezeigt, dabei steht der Datentyp, der die Knoten bildet, außen und der Datentyp, der die Kanten bildet, innen. Bei der Projektion, in der die Literatur die Knoten und die Objekte die Kanten bilden, sieht dies folgendermaßen aus:

[27]Die Namen sind bewusst kleingeschrieben, da dies der Benennung der Tabellen in der folgenden Anwendung entspricht. Das gilt für alle folgenden Angaben dieser Art.

literatur-objekt-literatur

Die zweite Zeile zeigt die Menge der Knoten und Kanten an. Dabei werden unverbundene Knoten mitgezählt.
Die dritte Zeile mit der Beschriftung “Maximum Density” gibt an, wie hoch die Dichte ist. Diese Zahl ist für gewöhnlich sehr klein.[28]
Die vierte Zeile zeigt die Menge der “Connected Nodes”, also aller Knoten die mindestens eine Kante haben. Auf Grundlage der Verbundenen Knoten wird ein weiterer Dichtewert, “Connected Density” errechnet.
In der fünften Zeile werden die durchschnittlichen Gewichte und die durchschnittliche Uniquity für den ganzen Graphen angegeben. Diese sind beim Two-mode Netz immer mit 1.0 angegeben.
In der sechsten Zeile ist die Menge der Komponenten ohne unverbundene Knoten angegeben.
In der siebten Spalte wird die durchschnittliche Größe der Komponenten angegeben.
Danach folgt eine Auflistung der fünf größten Komponenten. Diese ist mit der Menge von Knoten in der Komponente und der durchschnittlichen Gradzahl pro Komponente angegeben. Dahinter wird vermerkt, ob eine Komponente komplett verbunden ist. Das hilft zu bewerten, ob eine Komponente für weitere Analysen interessant ist. Ist eine Komponente komplett verbunden oder ist ihr durchschnittlicher Grad sehr hoch, dann ist davon auszugehen, dass sie nicht sehr interessant für weitere Analysen ist.

Technische Umsetzung

In der Implementation wird Schritt 3 paralell ausgeführt. Das ermöglicht die serverseitige Ausnutzung von Mehrkernsystemen. Das Vorgehen kann auch mit einigen Anpassungen in den Schritten 4,5,6 als Map-Reduce-Algorithmus[67] implementiert werden. Dadurch könnten auch massiv parallele Rechenressourcen auf Rechenclustern genutzt werden.
Im Map-Teil eines solchen Algorithmus werden viele parallele Teilprogramme ausgeführt, diese Map-Prozesse sind voneinander unabhängig. Im Reduce-Teil werden die Ergebnisse der parallelen Verarbeitung zusammengeführt.

Erweiterbarkeit

Der vorgestellte Algorithmus lässt sich in Schritt (3) und (5) sehr gut modifizieren, um das resultierende Netz differenzierter zu Formen. Mit zusätzlicher Information an Knoten und Kanten könnten dort Regeln eingebracht werden welche das Erstellen von Kanten genauer steuern. Das würde zu weniger Kanten führen. Hier könnte auch eine Art Filter implementiert werden, um Relationen, die viele leichte Kanten erstellen, zu vermeiden. Das ist im implementierten Algorithmus auch schon Vorgesehen, so wird verhindert, dass

[28] Dabei wird sich in allen Fällen, auch im bipartiten Netz auf die Dichte von ungerichteten unipartiten Netzen bezogen.

ein einzelner generell Verbundener Knoten nicht direkt zu einem vollständig verbundenen Graphen im Ergebnis führt.

3.6 Muster in der Arachne-Datenbank

Dieser Abschnitt beschäftigt sich mit Mustern die auftauchen, wenn verschiedenen Verbindungstypen aus der Arachne-Datenbank untersucht werden. Die Ergebnisse gehen hierbei aus dem Verfahren in Abschnitt 3.5 hervor. Die Arachne-Datenbank ist die Bild- und Objektdatenbank des Deutschen Archäologischen Instituts. Betrieben wird sie an der Universität zu Köln von der Arbeitsstelle für digitale Archäologie.

3.6.1 Geschichte und Datenbestand

Schon in der Frühzeit des Forschungsarchivs für antike Plastik in den 1970er Jahren gab es Experimente mit elektronischer Datenverarbeitung. Hierfür sollte erstmals mit dem GOLEM-System von Siemens ein abfragbarer, elektronischer Katalog erstellt werden. Das Projekt sollte in der Pilotphase etwa 2000 Bilder erfassen. Dieser Ansatz war jedoch zu umständlich und das Projekt schlug fehl.[64]
Die Datenbank, wie sie heute betrieben wird, stammt aus der Mitte der 90er Jahre. Sie wurde anfangs als Alternative zu einem überwuchernden Zettelkasten angelegt. So war die Datenbank anfangs eher ein Mittel zur Verwaltung der Fotobestände des Forschungsarchivs für antike Plastik. Erst Reinhard Förtsch erkannte in diesem digitalen Zettelkasten das Potenzial. Er machte in den Folgejahren die digitalisierten Bilder des Archivs und die dazugehörigen Objekt- und Kontextinformationen als Datenbank im Web verfügbar.[208]
Heute ist die Arachne-Datenbank eine der erfolgreiche Bild- und Objektdatenbanken im Bereich der Archäologie.
In ihrer langjährigen Geschichte ist die Arachne-Datenbank immer weiter gewachsen. Mit der Zeit entwickelte sich die Datenbank so weiter, dass eine ganze vielzahl von Kontexten eingepflegt werden mussten. Viele dieser Kontexte waren Angaben zu Orten und Literaturreferenzen. Dazu kamen Bauwerke und Reproduktionen von archäologischen Stücken.[217]
Die folgenden Analysen und Ansichten beziehen sich auf die Software und Verwaltungsstrukturen der Arachnedatenbank in Version 3.

3.6.2 Ansichten und Relationen

Die Arachne-Datenbank ist durch die diversen dokumentierten archäologischen Strukturen eine sehr heterogene Datenbank. Dies kann u.a. als Folge der Wiederverwendung von Topographien, Orte, Stilrichtungen, Sammlungsinformationen und vieler weiterer Arten von Entitäten gesehen werden.
Diese Diversität der Arachne machte es früh nötig, ein einfach zu erweiterndes Relationssystem zu erstellen. Zu diesem Zweck wurde Anfang 2006 ein dynamisch erweiterbares Relationssystem mit entsprechendem Kontextbrowser entwickelt. Dieses System legte fest, wie die einzelnen Datenbestände verknüpft sein können. Realisiert wurde es durch ein zentrales Register der Refenztabellen, eine Metastruktur, welche die Relationstabellen in

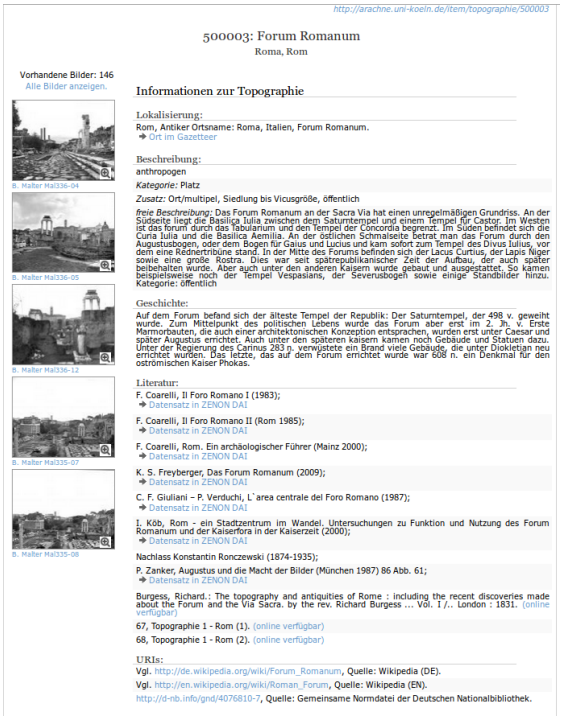
der Datenbank abstrakt beschrieb. So wurde sicher gestellt, dass Relationen sehr einfach nachgepflegt werden können. Jede dieser Relationen beschreibt theoretisch einen eigenen Graphen.

Ein Relationssystem wie in Arachne ermöglicht es, automatisiert einen multipartiten Graphen aus der Datenbankstruktur zu extrahieren. Dabei sind die Relationen nicht strikt typisiert.
In diesem System wurden dabei nicht nur die Verbindungen verwaltet, sondern es wurden auch Spezialinformationen, die zu einer Verbindung gehören, hinterlegt. Ein Stück kann beispielsweise einer Sammlung zugeordnet werden. Dabei hat diese spezielle Relation die Information, welche Inventarnummer das Stück in der referenzierten Sammlung hat. Eine spätere Version erweiterte das System dabei um eine Möglichkeit, Relationen mit Semantik zu versehen und Links zu externen Datenquellen zu setzen.
Die einzelnen Datensätze werden mit einer einfachen Ansicht von Datenbank-Einträgen mit einer Übersicht von Bildern dargestellt. Die Textfelder in der Datenbank sind sehr spezifisch. Viele Informationen, die erfasst werden können, treffen nicht auf alle Objekte zu. Daher erscheinen die Datensätze aus einer Datenbankperspektive oft "unvollständig". Die Ansicht der textbasierten Daten folgt daher in einer klassischen HTML Seite. Diese Datensatzansichten sind in Abbildung 3.16a[29] und 3.16c[30] abgebildet. Um diese Kontextrelationen darstellen zu können und Alternativen zu einer Listendarstellung zu bieten, wurde ein grafischer Kontextbrowser entwickelt. Die Knoten in dieser Ansicht verlinken auf die Datensätze. Der Kontext wurde dabei durch die Entitäten in der Datenbank definiert, die ein bis zwei Relationsschritte entfernt lagen. Die Knoten wurden auf konzentrischen Kreisen angeordnet und die Struktur des Kontexts wurde auf einen Baum reduziert. Diese Kontextbrowseransichten sind in Abbildung 3.16b und 3.16d zu sehen. Es ist klar ersichtlich, dass die einzelnen Datensätze unterschiedlich stark verknüpft sind. Aus netztheoretischer Sicht lässt sich, in Anlehnung an die Dichte eines Netzes, von verschiedenen Kontextdichten sprechen. Das Forum Romanum, welches eines der zentralen Monumente des Römischen Reiches ist, hat einen sehr viel größeren Kontext als das Relieffragment des Hermes.
Um die Funktionalität zu gewährleisten, mussten dafür sehr zahlreiche Kontexte des gleichen Typs in einzelne Knoten zusammengefasst werden. Dies ist insofern praktisch, da Kontexte sehr kleinteilig werden können. So kann eine Topografie viele Bauwerk umfassen, diese Bauwerke können Bauteile umfassen, welche wiederum Objekte beinhalten. Die Granularität ist hier stark von der Erfassungsgenauigkeit abhängig. Objekte können auch direkt an Bauwerke assoziiert werden ohne Bauwerksteile.
Ein anderes Beispiel sind Stiche in Büchern, die bestimmte Topografien oder Gebäude beschreiben. Das ist in Abbildung 3.16b unten links sichtbar. In der gleichen Abbildung werden der Palatin und das Kapitol aufgeführt, welche beide mit vielen Objekten ver-

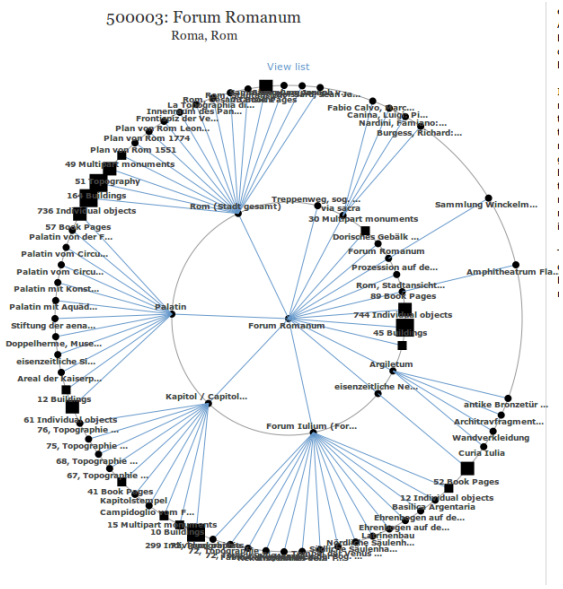
[29] `http://arachne.uni-koeln.de/item/topographie/500003`[100]

[30] `http://arachne.uni-koeln.de/item/objekt/1269`[100]

Abbildung 3.16: *Arachne Datensätze und Kontextbrowser*

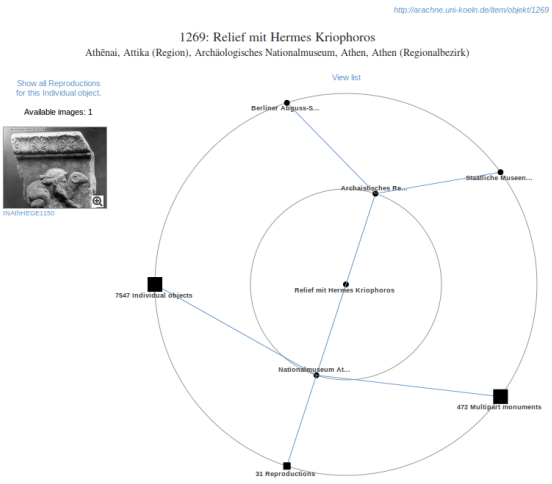


(a) Datensatzansicht Forum Romanum



(b) Kontextbrowser Forum Romanum

(c) Datensatzansicht Relief mit Hermes Kriophoros



(d) Kontextbrowser Relief mit Hermes Kriophoros

knüpft sind. Hier ist eine weitere Informationsreduktion zu sehen, viele Entitäten des gleichen Typs auf gleicher Position werden in quadratischen Knoten zusammengefasst. Diese Knoten referenzieren auf die Menge der Entitäten, die sie zusammenfassen.

Weitergehende Relationen

Gerade die gewachsene Struktur der Arachne-Datenbank wirft die Frage auf, ob sich aus dem Zusammenspiel der Daten Synergien ergeben die im Netz "versteckt" liegen.
Im Folgenden werden alle Verbindungen zwischen Entitäten-Klassen in der Arachne-Datenbank auf Muster in den Netzen untersucht. Die Verbindungen der Arachne Entitäten-Klassen werden meist über Referenztabellen abgebildet, es gibt jedoch auch Verbindungen mit geringerer Kardinalität. Die meisten Verbindungen können daher als bipartite Netze untersucht werden. Diese bipartiten Netze werden im Folgenden auf Grundlage ihrer Projektionen untersucht. Es werden alle Kombinationen von Datensätzen untersucht. Diese Voranalyse bietet dabei erste Hinweise auf eine mögliche interessante Struktur in den Relationen zwischen den Daten. Im Zuge dessen wurden alle Netze im Hinblick auf ihre Komponenten untersucht sowie auf den durchschnittlichen Grad der Knoten in den Komponenten. Diese Werte sind kostengünstig zu berechnen. Die Projektionen von bipartiten Netzen können in ihrer Struktur hinweise darauf liefern, ob die Daten interessante Netzstrukturen aufweisen. Auf der Grundlage dieser Kennzahlen kann die schiere Menge aus Kombinationen eingegrenzt werden, denn nicht aus jeder Zuordnung muss ein sinnvolles oder verbundenes Netz entstehen.

3.6.3 Buchseiten und Bauwerke

In diesem Abschnitt werden einzelne Buchseiten und Bauwerke in Beziehung gesetzt. In der automatischen Extraktion kamen große Strukturen zum Vorschein. Dies ist die Ausgabe des bipartiten Netzwerks:

Listing 3.3: *Analyseergebnisse Bauwerke zu Buchseiten*

```
bauwerk<-> buchseite
Nodes: 818232  Edges: 4386
Maximum Density: 1.3102261E-8
Connected Nodes: 4610 Connected Density: 4.1284878E-4
Avarage Edge Weight: 1.0 Avarage Edge Uniquity: 1.0
Number of Components: 262
Avg. size of Components : 17.595419
Largest Components :
1. Component size 1490 Avg. component Degree : 2.0
2. Component size 838 Avg. component Degree : 1.9976133
3. Component size 563 Avg. component Degree : 2.1030195
4. Component size 194 Avg. component Degree : 1.9896908
5. Component size 59 Avg. component Degree : 2.0
```

Die hier gezeigten Werte sind für bipartite Netze nicht aussagekräftig wie für Projektionen in unipartite Netze. Hier wird deutlich, dass der extrahierte Graph nur einen Bruchteil der Bauwerke und Buchseiten verbindet. Somit gibt es sehr viele Datensätze in der Datenbank über Buchseiten und Bauwerke, die keine Verbindung untereinander haben. Es wird in dieser Darstellung also nur ein Bruchteil der Daten betrachtet. Das Listing 3.3 zeigt weiterhin eine Auflistung der fünf größten Komponenten. Diese variieren um durchschnittlich zwei Verbindungen pro Knoten. In der Betrachtung der Projektionen des bipartiten Netzes, also dem unipartiten Netz, ergibt sich ein anderes Bild:

Listing 3.4: *Analyse Projektion Bauwerke über Buchseiten*

```
bauwerk-buchseite-bauwerk
Nodes: 8585  Edges: 276
Maximum Density: 7.49047E-6
Connected Nodes: 146 Connected Density: 0.026074633
Avarage Edge Weight: 1.0797101 Avarage Edge Uniquity: 0.4420287
Number of Components: 35
Avg. size of Components : 4.1714287
Largest Components :
1. Component size 58 Avg. component Degree : 6.6896553
2. Component size 8 Avg. component Degree : 4.25
3. Component size 5 Avg. component Degree : 4.0 Completely Connected
4. Component size 3 Avg. component Degree : 2.0 Completely Connected
5. Component size 3 Avg. component Degree : 2.0 Completely Connected
```

In der Projektion, in der die Bauwerke die Knoten bilden und die Buchseiten die Kanten, gibt es 35 Komponenten, die aus mehr als einem Knoten bestehen, siehe Listing 3.4. Dabei hat schon die viertgrößte Komponente nur drei Knoten. Auffällig ist dabei, dass in diesem Graphen nur 146 von 8585 Knoten überhaupt in einer Komponente von mindestens zwei Knoten bestehen. Die Ordnung der Komponenten in der Projektion entspricht dabei nicht der Ordnung im Ausgangsgraphen.
Die Analyse der Komponenten zeigt eine größte Komponente, die 58 Bauwerke umfasst. Der durchschnittliche Grad eines Knotens in dieser Komponente ist 6,69. Die Komponente ist interessant, da sie nicht komplett verbunden ist. In 3ten, 4ten und 5ten Komponenten sind alle Knoten mit allen anderen Knoten verbunden (hier mit “Completely Connected” beschriftet). Dies ist ein recht gewöhnliches Phänomen, es bedeutet jedoch, dass sich hier wahrscheinlich keine erkenntnisreichen Informationen im Sinne der Netzwerktheorie finden lassen.

Listing 3.5: *Analyse Projektion Buchseiten über Bauwerke*

```
buchseite-bauwerk-buchseite
Nodes: 809647  Edges: 975932
Maximum Density: 2.9775474E-6
Connected Nodes: 4155 Connected Density: 0.113086835
Avarage Edge Weight: 1.0000236 Avarage Edge Uniquity: 0.0022003986
Number of Components: 181
```

Abbildung 3.17: *Die größte Komponente des Bauwerk-Buchseiten-Graphen. Die blauen Knoten repräsentieren Buchseiten, die lila Knoten Bauwerke.*

(a) *Bipartiter Graph*

(b) *Das größte Cluster des Bauwerk-Buchseiten Graphen als Projektion über die Buchseiten. Übrig bleiben lediglich zwei Bauwerke.*

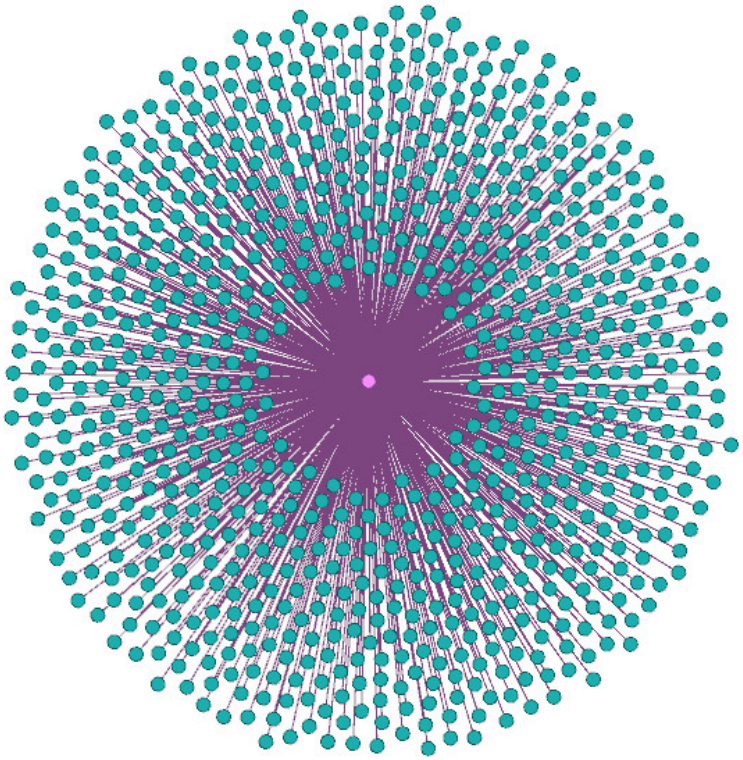
```
Avg. size of Components : 22.955801
Largest Components :
1. Component size 1488 Avg. component Degree : 802.0363
2. Component size 837 Avg. component Degree : 836.0 Completely Connected
3. Component size 505 Avg. component Degree : 26.237623
4. Component size 192 Avg. component Degree : 183.20833
5. Component size 51 Avg. component Degree : 11.960784
```

In der Projektion der Buchseiten über Bauwerke bestehen auch Cluster, die sehr dicht sind, diese sind jedoch oft nicht komplett verbunden.
Als Nächstes werden zwei Komponenten des Bauwerk-Buchseiten-Graphen genauer betrachtet. Als Erstes wird die größte Komponente beleuchtet. Sie enthält 1488 Buchseiten.

Wie Abbildung 3.17a zeigt, sind diese Knoten jedoch sehr ungleich verteilt. Es gibt nur zwei Bauwerke, aber eine große Menge von Buchseiten, die die Bauwerke beschreiben. Zwischen den Bauwerken sind zwei Buchseitenknoten verlinkt. Diese Buchseiten stellen die beiden Gebäude dar und schaffen so eine Relation, die auch in der Projektion der Komponente bestehen bleibt. Bei der Untersuchung dieser Komponenten mit einer Projektion in ein unipartites Netz, verschieben sich die Ergebnisse.
Das größte Cluster in der bipartiter Betrachtung zeigt dabei, dass es in der Projektion über die Buchseiten nur eine einzige Kante als Ergebnis hat. Diese Projektion in Abbildung 3.17b verdeutlicht, dass die beiden Bauwerke gemeinsam abgebildet sind. Für die Projektion über die Buchseiten sind nur zwei Seiten dieser Komponente von Bedeutung. Die anderen Einzeldarstellungen bleiben in den so entstehenden Projektionsdaten unbetrachtet. Bei den beiden Gebäuden handelt es sich um die Gebäude "Vatikanischer Papstpalast"[31]

[31] `http://arachne.uni-koeln.de/entity/11535`[100]

Abbildung 3.18: *Darstellung der zweitgrößten bipartiten Komponente des Bauwerk-Buchseiten Graphen.*



und "Museo Pio Clementino"[32]. Diese Bauwerke sind beide auf aufeinanderfolgenden Seiten des gescannten Buches "Georg Lippold. Die Skulpturen des Vaticanischen Museums. Band III,2 Text Berlin, 1956." [33]. Dort wurden die genannten Gebäude auf die Buchseiten verlinkt. Die Struktur ist dabei der Abfolge der Seiten und der linearen Beschreibung der Gebäudeteile geschuldet. Auf einer Seite wird von der Beschreibung des ersten Gebäudes auf die Beschreibung des zweiten Gebäudes oder dessen Inhalt eingegangen.
Das zweite Cluster besteht aus nur einem Bauwerk und vielen Buchseiten, die dieses Bauwerk beschreiben, siehe Abbildung 3.18. Diese bilden ein sternförmiges Pattern. Diese Struktur geht bei einer ungewichteten Projektion verloren. Es gibt jedoch Vorgehen, bei denen sich diese Informationen im Knotengewicht abbilden lassen. Interessanter ist das dritte Cluster. Es wird in Abbildung 3.19 dargestellt und der dazugehörige bipartite Graph wird in Abbildung 3.19a gezeigt. Die blauen Knoten sind Buchseiten, die lila Knoten sind Bauwerke.
Die Abbildung 3.19b zeigt den dazugehörigen unipartiten Graphen über die Buchseiten. Hier wird im Gegensatz zu Abbildung 3.17 deutlich, dass sich aus kleineren Komponenten von bipartite Graphen, wichtige Informationen herauslesen lassen. Das Netz zeigt mehrere kleine Cluster von Bauwerken und auch ein ganzes Netz von Bauwerken. Diese Komponente im Speziellen zeigt dabei Bauwerke aus Rom. Auf den Buchseiten werden sie in

[32] `http://arachne.uni-koeln.de/entity/10593`[100]

[33] `http://arachne.uni-koeln.de/item/buch/979`[100]

Abbildung 3.19: *Die drittgrößte Komponente des Bauwerk-Buchseiten-Graphen.*

(a) *bipartiter Graph*

(b) *Unipartiter Graph*

verschiedenen Darstellungen abgebildet.
In diesem Graphen vermischen sich dadurch einige Informationstypen. Dazu gehört der ortsgebundene Zusammenhang, der durch Ansichten dargestellt wird, wie Ansichten, die zwei Gebäude auf einem Bild abbilden und dadurch in Zusammenhang zueinander stehen. Es gibt Abbildungen von Querschnitten, die Vergleiche zwischen Gebäuden anstellen[34], Darstellungen die zwei Gebäude in einer perspektive einfangen [35] und ganze Ansichten, die mehrere Gebäude in einen gemeinsamen perspektivischen Kontext rücken [36].
Die zentralen Buchseitenknoten, die verschiedene Bauwerke auf einmal abbilden und eine hohe Betweenness Centrality aufweisen, sind stilisierte nicht geografische Karten Roms von Cavali [49].[37] Andererseits sind es aber auch rein referenzielle Abbildungsseiten. Hier sind mehrere Bauwerke auf einer Seite. Diese Seite beinhaltet verschiedene Abbildungen. Hier wird eher ein loser Zusammenhang hergestellt.[38] In einer Analyse der Zentralität der Buchseiten wie sie als grüne Knoten in Abbildung 3.19a dargestellt sind, lässt sich in Erfahrung bringen, welche Arten von Buchseiten eine allgemeine zentrale Lage einnehmen. Bei den höchsten Betweenness-Centrality Werten sind Karten zu sehen. Nach den Karten kommen in der Rangfolge die Abbildungen und dahinter die Ansichten. Dies lässt sich dadurch erklären, dass Karten mehrere Objekte auch außerhalb der Sichtdistanz abbilden, während Ansichten, stilisiert oder nicht, wahrscheinlich nähere Gebäude abbilden. Hierbei muss beachtet werden, dass die Daten nicht erschöpfend sind. Selbst wenn die Daten es für die Datenbank wären, gibt es abgebildete Dinge, die nicht in der Datenbank sind. Aufgrund der Popularität der Bauwerke und da die Buch-Digitalisate jüngerer Natur als viele Bauwerke sind, ist es wahrscheinlicher, dass die Bücher nach dem Einlesen in die Datenbank auf die Gebäude verlinkt wurden, die populär und allgemein bekannt sind. Dieses Problem wird in Kapitel 4 erläutert.
Ein Problem an der Vergleichbarkeit der Resultate ist, dass es mehrere Abbildungen auf einer Seite gibt und die Links nur auf die Seite, jedoch nicht auf ein Bild auf einer Seite verweisen. Die Seite stellt damit eine vorgegebene Informationsstückelung dar. Relationen werden dadurch teilweise willkürlich. Hier könnte von einem Rauschen auf der Granularitätsebene der Buchseiten gesprochen werden.
Durch die Projektion in ein unipartites Netzwerk verliert ein Netz an Informationen. Diese können jedoch in den Kantengewichten abgebildet werden.
Dieses Beispiel sollte beleuchten, dass die einfache Analyse eines bipartite Graphen hinsichtlich seiner Knotenanzahl nur eingeschränkt zeigen kann, was wirklich interessant ist, siehe Abbildung 3.3. Große Strukturen, hier Komponenten, können auch Ansammlungen von Links auf einen einzigen Knoten sein. Dies ist in einer auf Verbindungen fokussierten

[34] `http://arachne.uni-koeln.de/item/buchseite/274168`[100]

[35] `http://arachne.uni-koeln.de/item/buchseite/255268`[100]

[36] `http://arachne.uni-koeln.de/item/buchseite/215437`[100]

[37] Beispielsweise: `http://arachne.uni-koeln.de/item/buchseite/203741`[100], `http://arachne.uni-koeln.de/item/buchseite/203731`[100], `http://arachne.uni-koeln.de/item/buchseite/203739`[100]

[38] `http://arachne.uni-koeln.de/item/buchseite/307236`[100], `http://arachne.uni-koeln.de/item/buchseite/306987`[100]

Ansicht, wie der Graphentheorie, eher uninteressant und es gibt einfachere Methoden, Datensätze mit vielen Verknüpfungen zu suchen. Des Weiteren sind, wie in Abbildung 3.17b gezeigt, viele Knoten für die Entdeckung weiterer Strukturen überflüssig. Im größten Cluster brachten nur zwei Knoten eine parallele Darstellung mit sich.
In der praktischen Anwendung bedeutet das für einen Bearbeiter, dass er mithilfe dieser Informationen Datensätze so aufbereiten kann, dass alle Buchseiten, die Bauwerke in einem örtlichen Kontext zeigen, auch mit den entsprechenden Orten verbunden werden können.
Ein weiterer Anwendungsbereich dieser Daten ergibt sich aus kunsthistorischer Herangehensweise. Die Buchseiten, die hier verbunden sind, enthalten oft Abbildungen der Bauwerke. Dies macht einen Vergleich von Zuständen möglich, in denen Gebäude abgebildet wurden. Dies lässt Rückschlüsse auf Rekonstruktionen zu. Die digitalisierten Bücher sind außerdem alle aus dem Copyright gefallen, was den öffentlichen Zugang zu digitalen Scans erleichtert. Analysen wie diese wurden schon von Schich an der CENSUS-Datenbank vorgenommen.[219]

3.6.4 Objekte und Sammlungen

In der Arachne-Datenbank gibt es eine Zuordnung von Objekten zu Sammlungen. Der Datensatz umfasst bestehende sowie aufgelöste Sammlungen. Objekte können also mehreren Sammlungen zugeordnet sein. Dabei ist auf Grundlage der Datenbank aber nicht direkt ersichtlich, wann welches Objekt in welcher Sammlung war. Diese Information ist nicht explizit codiert. Vorstellbar ist aber eine Ableitung dieser Information zu den Sammlungen.
In der bipartiten Betrachtung wird festgestellt, dass es sich hierbei um einen sehr großen bipartite Graphen handelt.

Listing 3.6: *Analyseergebnisse Objekte zu Sammlungen*

```
objekt<-> sammlungen
Nodes: 158176  Edges: 86246
Maximum Density: 6.894306E-6
Connected Nodes: 73305 Connected Density: 3.2100197E-5
Avarage Edge Weight: 1.0 Avarage Edge Uniquity: 1.0
Number of Components: 418
Avg. size of Components : 175.37082
Largest Components :
1. Component size 52888 Avg. component Degree : 2.481924
2. Component size 6311 Avg. component Degree : 2.0009508
3. Component size 908 Avg. component Degree : 1.9977974
4. Component size 871 Avg. component Degree : 3.097589
5. Component size 751 Avg. component Degree : 1.9973369
```

Die größte Komponente hat eine Größe von 52888 Knoten. Die zugehörige GraphML-Datei hat eine Größe von 23 Megabyte.
Aufgrund der Größe und Menge der Objekte werden wir in diesem Fall die Projektion von

Sammlungen über Objekte genauer betrachten.
Die Projektion von Objekten über Sammlungen erzeugt einen stark verbundenen Graphen. Die verbundenen Knoten haben eine Dichte von über 0,009.

Listing 3.7: *Analyse Projektion Objekte über Sammlungen*

```
objekt-sammlungen-objekt
Nodes: 156924  Edges: 4206187
Maximum Density: 3.4161948E-4
Connected Nodes: 29358 Connected Density: 0.009760688
Avarage Edge Weight: 1.0213503 Avarage Edge Uniquity: 0.0037726224
Number of Components: 373
Avg. size of Components : 78.70777
Largest Components :
1. Component size 10583 Avg. component Degree : 233.2934
2. Component size 1396 Avg. component Degree : 511.66763
3. Component size 907 Avg. component Degree : 906.0 Completely Connected
4. Component size 866 Avg. component Degree : 865.0 Completely Connected
5. Component size 558 Avg. component Degree : 557.0 Completely Connected
```

Sammlungen über Objekte

Die Projektion der Sammlungen über Objekte legt dabei einen Zusammenhang von Sammlungen untereinander nahe. Ein Objekt wurde einmal in der einen und dann in der anderen Sammlung registriert. Daraus ergibt sich eine Art von Austauschnetz.
Doch ist bei der Recherche in der Datenbank ersichtlich, dass die Sammlungen, die in der Datenbank angegeben sind, oft nicht mehr existieren. Insofern haben Objekte oft durch die Auflösung einer Sammlung ihren Weg in andere Sammlungen gefunden. Auf diese Weise bleiben oft Gruppen von antiken Statuen eines "Themas" zusammen.
Das Netz könnte also auch als eine Art Flussnetz gesehen werden. Dabei ist die Idee des Flusses der Objekte über Sammlungen hinweg interessant. Das Cooccurence-Verfahren zu den Sammlungen hat hier nur einen Nachteil: Es verbindet alle Sammlungen mit allen anderen, in denen die gleichen Objekte vorkamen. Das heißt auch, dass der erste und letzte Aufbewahrungsort eines Objektes verbunden werden. Die transitive Beziehung wird explizit abgebildet. Für ein Handels- oder Flussnetz ist diese Information leider fehlerhaft, da nach direkten Einflüssen und Kontakten gesucht werden könnte. Es kann daher nur eingeschränkt von einem Austausch, Fluss oder Handelsnetz gesprochen werden.

Listing 3.8: *Analyse Projektion Sammlungen über Objekte*

```
sammlungen-objekt-sammlungen
Nodes: 1252  Edges: 1036
Maximum Density: 0.0013229033
Connected Nodes: 514 Connected Density: 0.00785795
Avarage Edge Weight: 15.791506 Avarage Edge Uniquity: 12.194961
Number of Components: 33
```

```
Avg. size of Components : 15.575758
Largest Components :
1. Component size 434 Avg. component Degree : 4.525346
2. Component size 10 Avg. component Degree : 2.4
3. Component size 5 Avg. component Degree : 2.8
4. Component size 3 Avg. component Degree : 1.3333334
5. Component size 2 Avg. component Degree : 1.0 Completely Connected
```

Visualisierung

Für die hier vorliegende Visualisierung in Abbildung 3.20 wurde der ForceAtlas2 Algorithmus von Gephi verwendet. Für die Gewichtung der Kanten bei der Positionierung wurde die durchschnittliche Uniquity genommen. Die Zuordnung von Objekten zu mehreren Sammlungen macht diesen Wert schwächer. Durch das Projektionsvorgehen haben sich, wie besprochen redundante Kanten gebildet. Diese werden durch die durchschnittliche Uniquity schwächer bewertet, haben daher weniger Einfluss auf die Positionierung. Für die Darstellung der Kanten wurde das einfache Gewicht aus dem Coocurence-Verfahren gewählt. In den Daten gibt es einige sehr stark repräsentierte Fragmente.
Der Bias ensteht durch die Auswahl der eingegebenen Daten und Geschichte der Datenbank. Dabei kann man die stärksten Kanten zwischen “Stosch” und “Staatliche Museen, Antikensammlung, Altes Museum Berlin” erkennen. Hier ging eine Sammlung von Gemmen im Gesamten in die Sammlung der staatlichen Museen über. Diese Gemmensammlung ist in der Arachne stark dokumentiert. Daher bildet sich hieraus die stärkste Kante in diesem Kontext. Wie in Abbildung 3.21a zu sehen ist, die Kante zwar die Stärkste, doch die Wichtigkeit der Sammlung Stosch ist in diesem Kontext so gering, dass das Label für diesen Knoten unter die Darstellungsgrenze fällt. Ähnliche Fragmente finden sich öfters. Dabei ist der Bias in der Gesamtdarstellung kaum zu erkennen.
Die Größe und Wichtigkeit wird hier durch die Betweenness Centrality ausgedrückt. Diese ist aufgrund des transitiven Charakters des Netzwerks angebracht. Außerdem kann so dem beschriebenen Bias entgegengewirkt werden und eine globalere Gewichtung angenommen werden. Die Farben der Knoten kommen aus dem Bondell Clustering Algorithmus wie in Gephi implementiert. Er zeigt dabei teilweise auch geographisch einordnbare Cluster. Das hellrote Cluster zeigt dabei augenscheinlich die englischen Sammlungen. Das hellblaue Cluster zeigt italienische und römische Sammlungen. Das gelbe Cluster hat als Mittelpunkt athenische Museen. Das violette Cluster bildet sich um die Sammlungen des Pergamon Museums herum. Neben einem Kategorisierungsaspekt soll die Einfärbung dem Leser der Visualisierung helfen, eine Position in Detailansichten wieder zu finden.

Nahbetrachtung

Beispiele für die Relation zwischen Sammlungen und Objekten lassen sich in der Datenbank recherchieren.
In den Datensätzen der Sammlungen sind teilweise Informationen zum Werdegang der Sammlungen verzeichnet. Hierzu gehören das Gründungsdatum und Auflösungsdaten so-

Abbildung 3.20: *Größten Komponente in der Projektion von Sammlungen über Objekte.*

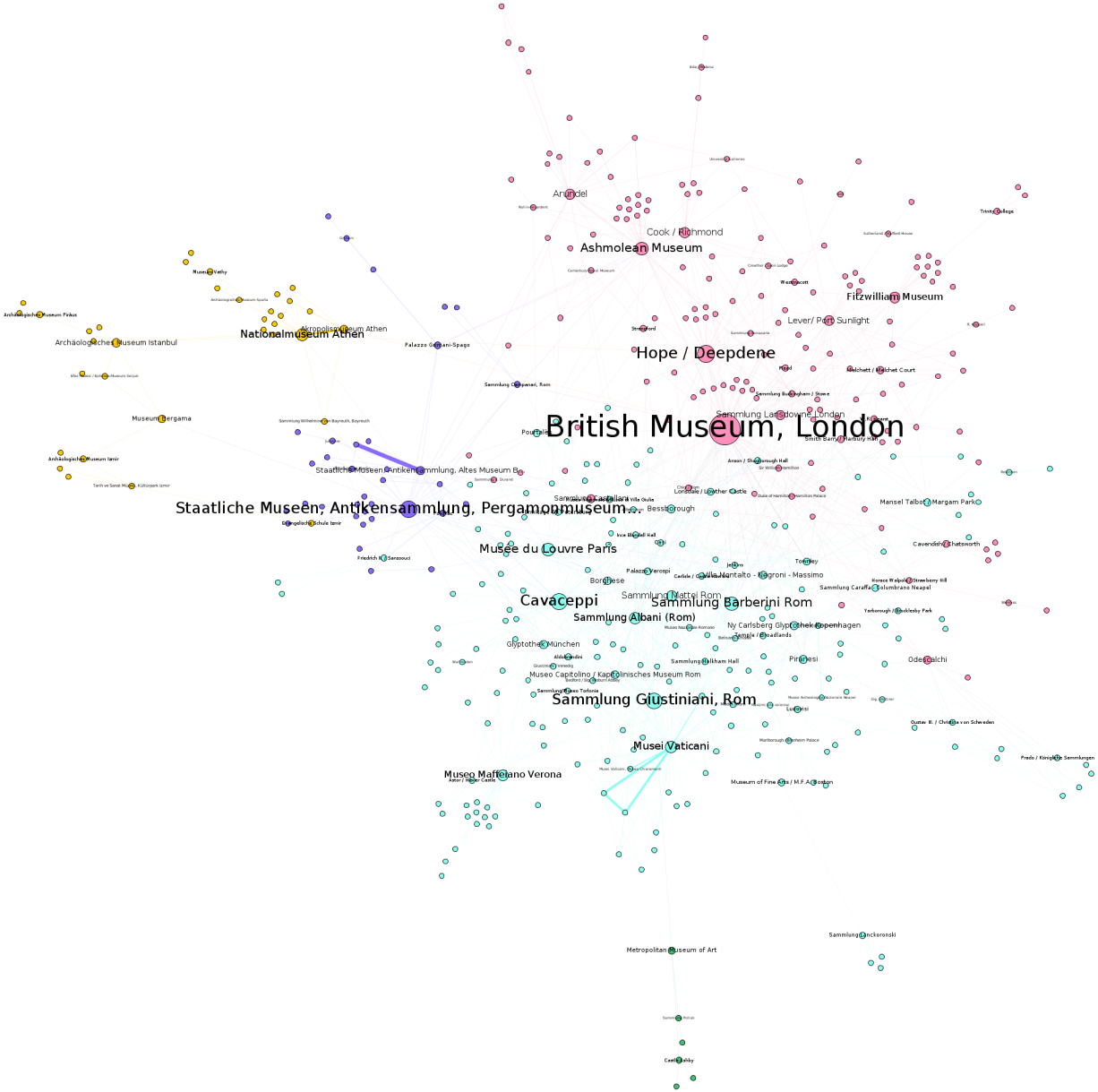
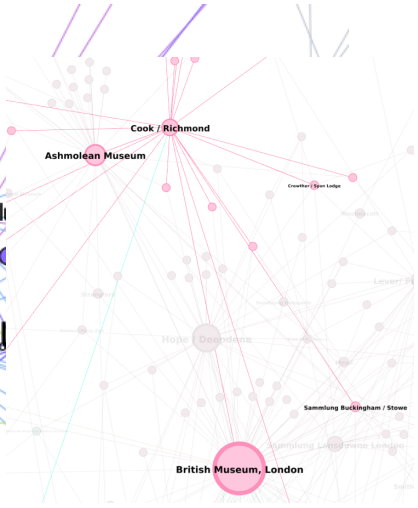


Abbildung 3.21: *Details zur Darstellung 3.20*

(a) *Detail mit der am höchsten gewichteten Kante zwischen der Sammlung Stosch und der staatlichen Antikensammlung Berlin.*



(b) *Auschnitt um das Ashmolean Museum*

wie die Umstände der Auflösung etc. Weitere Analysen werden dadurch erschwert, dass nicht alle Datensätze diese Information enthalten. Das hängt mit dem Eingabekontext der Daten zusammen. Es gab Projekte und Eingabekampangen die mehr Wert auf die Eingabe dieser Informationen legten als Andere. Projekte die Sammlungszentriert Arbeiten erfassen dabei Sammlungen meist vollständig. Es können jedoch auch gezielt Objektgruppen untersucht worden sein, für die dann eine "komplette" Sammlungsgeschichte angelegt worden ist. In diesem Fall wäre eine Vollständigkeit vom Objekt aus zu sehen. Andere Referenzen könnten aus Katalogen stammen etc. oder eingegeben worden sein, da die Informationen verfügbar waren. Dies würde eher zufällig verteilte Referenzangaben erstellen. Es ist jedenfalls nicht gegeben, dass alle Informationen bei allen Datensätzen komplett sind.

Nun wird auf ein paar Verbindungen zwischen den Sammlungen genauer eingegangen:

Ashmolean Museum in Oxford[39]

Cook / Richmond in Doughty House, Richmond / Surrey [40]

British Museum in London [41]

Dies ist in Abbildung 3.21b als Detailansicht mit Hervorhebung aus der Übersichtskarte 3.20 dargestellt.

Dabei ist zu bemerken, dass das Ashmolean Museum und das British Museum nicht verbunden sind. Die Geschichte hinter diesem Pfad ist, dass die Sammlung nach ihrer Auflösung verkauft wurde. Dies ist aus dem Datensatz zur Sammlung Cook / Richmond[42] ersichtlich. Dabei ist auch explizit festgehalten, dass die Skulpturen in die anderen Museen übergingen.

[39] `http://arachne.uni-koeln.de/entity/1240882`[100]

[40] `http://arachne.uni-koeln.de/entity/1240855`[100]

[41] `http://arachne.uni-koeln.de/entity/1241635`[100]

[42] `http://arachne.uni-koeln.de/entity/1240855`[100]

Die Sammlung Cook / Richmond stammt aus der Sammlung von Esq. Francis Cook. Dieser hat seine Stücke an seine Kinder vermacht, welche die Sammlung dann erweiterten.
Die Erkenntnis der Komplexität der Daten erfordert an dieser Stelle ein spezielles Analyseverfahren.
Interessanterweise lässt sich der Zusammenhang zwischen zwei Sammlungen nicht nur als Betrachtung der Verbindung durch die Weitergabe von Objekte sehen, sondern auch über die von Personen und deren Verwandtschaftsverhältnisse.

3.6.5 Literatur und Objekte

Die Arachne-Datenbank beinhaltet eine große Menge an Literaturreferenzen. Sie wurden im ursprünglichen Kontextbrowser nicht dargestellt. Eine archäologische Publikation hat für gewöhnlich einen Katalog im Anhang, anhand dessen argumentiert und beschrieben wird. Diese Daten lassen die Ähnlichkeit von Texten anhand von den in Ihnen beschriebenen Objekten auswerten. Mit einfachen Textanalysen ließen sich diese Informationen nicht ohne Weiteres extrahieren.[65]

Listing 3.9: *Analyse Projektion Literatur über Objekte*

```
literatur-objekt-literatur
Nodes: 18622  Edges: 118648
Maximum Density: 6.8432296E-4
Connected Nodes: 10821 Connected Density: 0.0020267295
Avarage Edge Weight: 1.82529 Avarage Edge Uniquity: 0.3716853
Number of Components: 270
Avg. size of Components : 40.077778
Largest Components :
1. Component size 10088 Avg. component Degree : 23.393736
2. Component size 36 Avg. component Degree : 5.3333335
3. Component size 6 Avg. component Degree : 2.0
4. Component size 6 Avg. component Degree : 2.6666667
5. Component size 6 Avg. component Degree : 3.6666667
```

Da die größte Komponente aus mehr als 10.000 Knoten besteht, ist eine sinnvolle Visualisierung nicht ohne Weiteres möglich. Dabei entsteht ein sogenannter Hairball, ein unentwirrbares Netz aus dem keine Information abgelesen werden können. Bei der Analyse dieser Daten können auch die Gewichte nur bedingt aussagekräftig.
Hier muss eine Priorisierung der Kanten vorgenommen werden, um den Graphen aus zu dünnen. Dafür wurde ein Filter über die Average Uniquity auf den Kanten verwendet. Bei den Versuchen, andere Werte zu verwenden, fiel auf, dass der Bias auf den Kanten sehr stark war, die Nennung in der Datenbank hing stark von allgemein bekannten Objekten ab. Dabei blieb das Problem, dass oft zitierte Objekt viel Gewicht auf einer Kante erzeugen können. Ein oft beschriebenes Objekt hat dabei eigentlich keine starke Aussage, da es keine einzigartige Beziehung zwischen Werten beschreibt.
Der Filter wurde auf alle Kanten, die einen Uniquity Wert von 0.5 und kleiner aufwiesen. Oft zitierte Objekte erzeugen einen sehr kleinen Uniquity Wert. Die aus diesen häufig

Tabelle 3.3: *Korrelation nach Pearson zwischen den berechneten Kantengewichten im ungefilterten Graphen der Projektion der Literatur über Objekte*

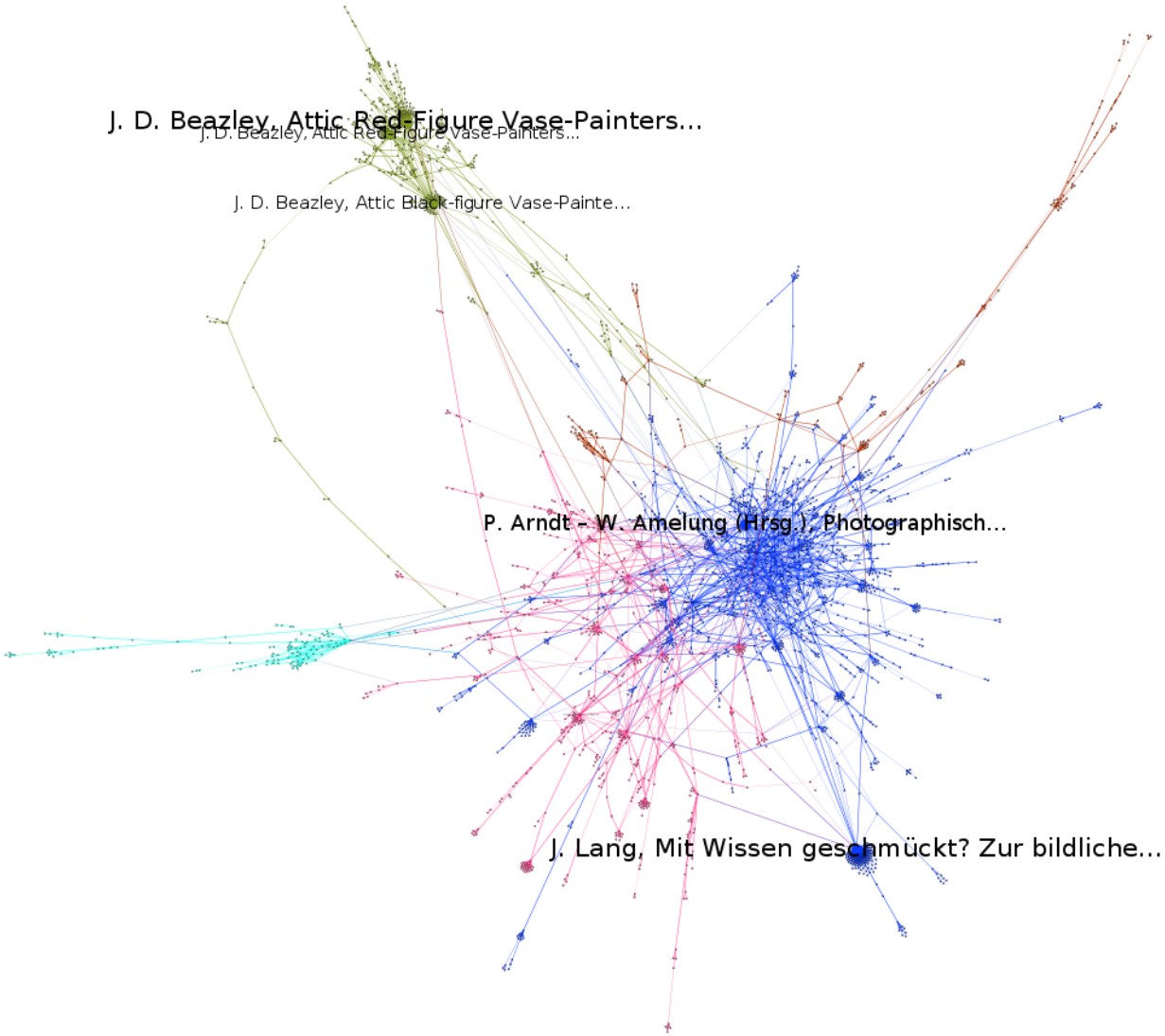
X	Uniquity	Av. Uniquity	Gewicht
Uniquity	1.0	0.07181864	0.93152163
Av. Uniquity	0.07181864	1.0	0.05360027
Gewicht	0.93152163	0.05360027	1.0

zitierten Objekten entstehenden Kanten werden durch den Filter entfernt.
Durch das Filtern wurde die Visualisierung in Abbildung 3.22 möglich. Die Labels und Knotengrößen wurden nach dem Grad der Knoten vergeben. Die Gruppen, die der Clustering-Algorithmus gefunden hat, können bei Betrachtung nur grob benannt werden.
Oben links, etwas abseits liegt das Grünbraune Cluster, dies sind Veröffentlichungen von Beazley zu rotfigurigen Vasen. Am unteren Ende findet sich die Arbeit eines der ehemaligen Mitarbeiter der Arachne-Datenbank, Jörn Lang. Das blaue Cluster in der Mitte der Abbildung stellt nach wie vor einen Hairball dar.

3.6.6 Fazit

Die Arachne-Datenbank ist eine Sammlung von teilweise sehr heterogenen Daten. Diese entstanden aus vielen verschiedenen Quellen und im Rahmen diverser Projekte. Es wurde dargestellt, dass sich Erkenntnisse aus den Projektionen und Visualisierungen abgelesen werden konnten. Die Maßzahlen haben gezeigt, wo relevante Netze zu finden sind und ob sie eine Dichte aufwiesen, die von Interesse ist. Die Netze waren nicht zu dicht und nicht zu sehr fragmentiert, um eine Information zu transportieren.
Im Falle der Betrachtung von Bauwerken und Buchseiten konnte gezeigt werden, dass sich Emergenzen ergeben können, die es beispielsweise Kunstgeschichtlern erleichtern können, verschiedene Darstellungen von in Relation gesetzten Bauwerken und gemeinsame Darstellungen zu finden. Aufgrund der Art der gewählten Subjekte war es nur logisch, dass sich ein örtlicher Rahmen um die Bauwerke schließt. Die hier betrachteten Fehler bezogen sich häufig auf die Granularitätsebene. Eine Buchseite konnte mehrere Abbildungen beinhalten. Für weitere Untersuchungen könnten solche unerwünschten Datenfragmente identifiziert und herausgefiltert werden. Diese Daten könnten auch Bearbeitern ermöglichen, Daten in der Datenbank mit Hilfe von Vorschlägen zu verfeinern.
Bei den Relationen zwischen Sammlungen und Objekten konnte eine Art Karte der Beziehungen zwischen Sammlungen erzeugt werden. Dabei wurde der Austausch von Objekten zum ordnenden Element. Mit einiger Arbeit und weiteren inhaltlichen Analysen lassen sich die Daten und Fehler minimieren und ein wirkliches Transaktionsnetz zwischen Sammlungen täte sich auf. Mithilfe dieses Netzes wäre es möglich, die Transitionen und Beziehungen von Sammlungen geschichtlich nachzuvollziehen. Dies ist vor allem durch das direkte Zurückgreifen auf die Quelldaten in der Datenbank durch in der Datenstruktur hinterlegten Links möglich.
Die Untersuchung der Literatur über die in ihnen beschriebenen Objekte nutzte eine der einzigartigen Eigenschaften der objektbasierten Literatur aus. Die konkrete Beziehung

Abbildung 3.22: *Bipartite Darstellung der zweitgrößten Komponente Literatur über Objekte*



und Betrachtung der gemeinsam zitierten Objekte, die eine gewisse Einzigartigkeit haben, lassen die Felder und Fokuspunkte der Archäologie aufleuchten. Dabei lassen sich Konzentrationen um Bestände beobachten, die sehr in der Datenbank präsent sind.
Die Arachne-Datenbank enthält viele Informationen über verschiedenste Objekte und Kontexte. Diese Informationen müssen konsistent gehalten werden.
Ein Vorteil einer solchen Datenstruktur liegt darin, dass der Kontext einer Information sehr breit aufgestellt sein kann. Es können zu vielen Dingen aus dem Umfeld der archäologischen Objekte Bezüge hergestellt werden. Diese Bezugspunkte müssen vorhanden sein, um referenziert werden zu können. Das ist eine Grundvoraussetzung, damit sich Emergenzen ergeben, die in sehr heterogenen Datenbeständen nicht auftreten können.
Der Nachteil ist die Breite und Heterogenität der Information. Daten aus verschiedenen Projekten mit verschiedenen Schwerpunkten erzeugen komplexe Strukturen. Komplexe Strukturen bergen eine starke Anfälligkeit für Probleme.
Es wäre also besser, wenn es kleinere Fachdatenbanken gäbe. Das ist im Kontext der Arachne-Datenbank passiert. Die Ortsverwaltung wurde in den iDAI.Gazetteer[221, 1] ausgelagert, um ein wiederverwertbares Ortsregister zu erschaffen. Zusammen mit der Literaturverwaltung iDAI.bibliography / Zenon[2], des DAI, hat sich hier eine spezialisierende Aufteilung der Arachne bewirkt. Ein solches Vorgehen ist natürlich auch nicht unproblematisch, da "Datenhoheit" abgegeben wird. Doch ob eine Datenbank wirklich alles verwalten muss und kann, ist eine der wichtigeren Fragen in diesem Zusammenhang. Diese Frage wird auch im nächsten Kapitel besprochen. Dort wird die Frage nach einem generellen informationsübergreifenden Netz aus Daten gestellt.

3.7 Zusammenfassung

In diesem Kapitel wurden einige Standarddatentypen der Geisteswissenschaften und ihre Umwandlung in Netzwerke untersucht.

3.7.1 Problematische und offensichtliche Hubs

Die Beispiele mit den XML-Daten zeigten einfach Fragestellungen und Visualisierungen von Datensätzen. Diese Visualisierungen konnten in ihrer Form nur sehr simple Fragen beantworten und einen Überblick verschaffen. Hierzu muss in die Visualisierung eingegriffen werden. Dabei kann Offensichtliches, wie der Werksnamensgeber Macbeth oder das Ausblenden der Akte des Stückes gemeint sein. Dieses Problem gab es bei den Gesetzestexten in dieser Form nicht. Die Gesetzestexte konnten recht eindeutig identifiziert werden. In der Betrachtung der Wahldaten konnten diese Fragmente auch nicht auftauchen, da jede Position eine geringe Höchstanzahl an Kanten haben konnte. Diese bezog sich auf die Anzahl der Parteien, deren Ablehnung und Zustimmung als verbindende Größe fungierte. Eine Hub-Bildung hätte hier einen Fehler im Programmablauf sichtbar gemacht.

3.7.2 Beschriftung

Ein Problem zeigte sich auch durch die Beschriftung der Knoten. Diese Gesetze hatten lange Namen. Auch bei der Visualisierung der Arachne-Daten fiel dieses Problem auf. Auch die meisten Datensätze in der Arachne haben lange Namen.
Im Fall der Wahldaten ist dieses Problem nicht in diesem Ausmaß hervorgetreten. Hier gab es die Wahlbezirke, welche meist keine originären Namen hatten, da sie aufgrund anderer Kriterien als ihrer Bezeichnung zusammengefasst wurden. Ihre Identität wurde durch Farben erstellt. Dies hatte den Vorteil das eine visuelle "Identität" aufgrund einer anderen Eigenschaft gefunden wurde. Die Rückprojektion über die Parteipräferenzen war dabei sehr einfach zu lesen, da die Namen der Parteien für gewöhnlich abgekürzt werden und die Parteien gängige Kürzel verwenden. Sie sind für den politisch Interessierten einfach zu erfassen.
Symbole oder kurze Worte können an dieser Stelle hilfreich sein, um Überlappungen zu verhindern.

3.7.3 Ort und natürliche Ordnung

Die reale Position oder Ordnung und ihre thematische Darstellung wurde auch behandelt. Die Gesetzestexte wurden in der Ordnung nach Seitenzahlen im Kreis dargestellt und erhielten je nach ihrer Position eine farbliche Kennung. Die Wahlbezirke wurden mit einer Doppeldarstellung farblich codiert. Dies ließ Rückschlüsse auf das politische Spektrum und auf das geografische Gebiet zu.
In der Arachne zeigte sich der Ort als zusammenhaltendes Glied einer Komponente. Die Gebäudedarstellungen auf Buchseiten zeigten implizit einen geografischen Raum. Das war die Folge aus der Darstellung von Gebäuden miteinander, die sich zusammen auf Karten oder auf gemeinsamen Abbildungen befanden. Auch die Transaktionskarte aus der Betrachtung von Sammlungen und Objekten enthält eine starke geografische Komponente, da Objekte dem ersten Anschein nach gerade zwischen geografisch nahen Museen und Sammlungen getauscht wurden. Dies ließ sich auch aus der farblichen Codierung des Clusterings ablesen. Die Betrachtung der Literaturzitate und der Objekte lässt keinen so klaren Schluss zu. Doch lassen die Verbindungen der Sammlungen und sammlungszentrierten Publikation einen ähnlichen Schluss zu. Hier wäre das Filtern von geografisch sammlungszentrierten Publikationen interessant, um stark ortsunabhängige thematische Publikationen zu finden.

3.7.4 Grenzen und Entgrenzung

Die gezeigten Darstellungen sind an sich immer "natürlich" eingegrenzt. Es wurden einzelne Dateien und Datensätze untersucht und sowie in sich geschlossene Datenbanken.
Gesetzestexte sind nicht in sich geschlossen, sondern verweisen explizit auf andere Gesetzestexte aus der deutschen Rechtssprechung. Das wurde im Rahmen dieser Arbeit nicht untersucht. Doch ist es möglich, weitere Artikel aus anderen Rechtsbüchern zu erkennen. Die Darstellung der Wahlergebnisse ist theoretisch durch ähnliche Daten aus angrenzen-

den Stadtgebieten erweiterbar, so denn diese in der gleichen Form vorhanden sind. Dies ist insofern interessant, da eine Stadt niemals für sich selbst existiert, sondern Verbindungen zum Umland bestehen und so Metropolregionen entstehen. Die Verwaltungsgrenzen sind meist für die Personen, die ihren Wohnort wählen, nicht in dem Maße ausschlaggebend. Daher lassen sich so die Horizonte erweitern.
Eine Entgrenzung ist auch für die Arachne-Datenbank denkbar. Die erwähnte Aufteilung in diverse Spezialsysteme ist dabei ein möglicher Schritt in die Entgrenzung der Daten. Einen Schritt in richtung entgrenzung der Arachne-Datenbank machte das Pelagios Projekt, es wird in Abschnitt 4.2.3 genauer betrachtet.

Kapitel 4

Netze von Datennetzen

Im letzten Kapitel wurde die Verarbeitung von Netzen in kontrollierten Datenstrukturen beschrieben. Die Daten waren meist statisch und es bestand voller Zugriff auf eine feste, nicht netzbasierte Datenstruktur. Dabei wurden verschiedene Typen von Daten besprochen, genauso wie besprochen wurde, wie sich aus diesen Netze extrahieren lassen. Dieses Kapitel beschäftigt sich nun mit der Herausforderung und den Möglichkeiten von Daten, die verteilt abgelegt und als semantisches Netz ausgezeichnet sind. Datenhaltung findet in diesen Systemen nicht in geschlossenen Datenbanken statt. Das Ziel dabei ist, Daten besser teilbar[1], allgemein beschreib- und rezipierbar zu machen. Weiterhin ist das Netzparadigma klare Grundlage dieser Datenstruktur, wie im RDF-Standard beschrieben. Diese Datenstruktur ist der bis jetzt verwendeten Graphentheorie sehr ähnlich.
Im letzten Kapitel wurden einige Methoden vorgestellt, wie aus ursprünglich nicht relationalen Daten Graphen gemacht werden.
Dass die Bereitstellung der Daten als Netz nicht zwangsweise eine Analyse der Daten vereinfacht, sondern die schiere Menge der Daten und die Komplexität ihrer Beschreibung andere Problem mit sich bringt, wird im folgenden Kapitel erläutert.

4.1 Semantic Web und Linked Data

Das Semantic Web ist eine Vision von Tim Berners Lee, dem Erfinder des World Wide Web. Es sollte das maschinenlesbare Equivalent zum dokumentbasierten WWW werden. Die semantische Auszeichnung und die sich daraus herleitende automatische Verarbeitbarkeit sollten die Möglichkeiten der Automatisierung stark erweitern. Das Semantic Web, welches auch von Computern "verstanden" werden soll, wurde als Idealziel der künstlichen Intelligenz gesehen. Die Idee des Semantic Web wurde dabei durch das pragmatische Ziel von Linked Data eingeholt.
Das Semantic Web hatte die Vision, durch die Auszeichnung von Daten mit logischen Vokabularen, sogenannten Ontologien, Daten allgemein verständlich zu beschreiben. Daten in Datenbanken beinhalten nicht zwangsweise eine Dokumentation oder eine Struktur, die von mehr Personen als dem Systembetreiber verstanden werden können. Gemeingültige

[1] Hier geht es nicht um das Zerteilen, sondern das Verteilen von Daten; Engl.: "to share".

Ontologien sollen eine einheitliche, vergleichbare Weltsicht etablieren. Die Daten, die darin ausgezeichnet sind, wären daher einheitlich verarbeitbar und ihre Struktur somit für jeden verständlich.
Die Idee ging aus dem Problem hervor, dass Datenbanken je nach Anforderung modelliert werden und in sich geschlossen sind. Sie enthalten jedoch große Mengen von Wissen, das potenziell auch für Dritte interessant ist.
Das Prinzip der geschlossenen Datenbank ist für die konkrete Verarbeitung und den direkten Zweck, der mit der Datensammlung verfolgt wird, sinnvoll. Dieses Vorgehen hat jedoch zur Folge, dass einzelne Datenbanken untereinander inkompatibel sind.
Schlüssel sind in jeder Datenbank anders vergeben und damit auch nicht außerhalb ihres Kontexts verständlich und verwendbar.

4.1.1 Semantic Utopia

Im Zentrum der Semantic Web Idee steht die Verwendbarkeit einer kritischen Masse von verteilten Daten. Dabei sollte aus vorhandenem Wissen neues Wissen hergeleitet werden. Dies ist eine Vision, die mit der Tradition der künstlichen Intelligenz zusammenhängt.
In der Vision von Tim Berners Lee sollten künstliche Intelligenzen, sogenannte Agenten, autonom Aufgaben erfüllen können. Ein Beispiel für eine solche Aufgabe wäre das Aushandeln eines Zahnarzttermins.[25]
Um dieses Unterfangen zu realisieren, sind eine Vielzahl an Regeln nötig. Diese Regeln dienen zur Umformung des vorhandenen Wissens in neues oder handlungsrelevantes Wissen. Sogenannte Inferenz-Maschinen[2] leiten aus vorhandenen Daten Informationen ab. Diese umgeformten Daten würden von Agentenprogrammen[3] verwendet, um Aufgaben zu lösen und Fragen zu beantworten.
Die dafür nötigen Daten müssten potenziell von allen Agenten verarbeitet werden können. Das Sprechen der gleichen "Datensprache" ist ein zentraler Punkt dieser Vision. Dadurch wird die Kommunikation zwischen Agenten erst möglich.
Im Falle des Arzttermins muss beispielsweise ein Format vorliegen, in dem Termine ausgezeichnet werden können. Ein anderes System aus Regeln befähigt den Agenten Termindaten, mit anderen Kontexten, in eine gemeinsame Terminvereinbarung zu verwandeln.

Der implementierte Semantic Web-Stack

Die technologische Utopie des Semantic Web setzt einen Stapel von Technologien voraus. Diese Technologien sind notwendig, da ein automatisierter Austausch von Daten eine nicht triviale Aufgabe darstellt. Hier flossen viele wichtige Lektionen aus der Geschichte der Computersysteme zusammen.
Auf der untersten Ebene des Stacks wurde Unicode[250] als Zeichensystem gewählt, da sich fast alle bekannten Zeichen, inklusive antiker Sprachen, phonetischer Zeichen, Emoticons etc. abbilden lassen.

[2] englisch reasoner

[3] Bei Tim Berners Lee: agents.

Die allgemeine Identifikation von Dingen sollte über eindeutige Identifier, URIs, geregelt werden. Diese URIs sollten nachhaltig referenzierbar sein, daher nicht nach einiger Zeit berschwinden. Außerdem sollen sie über das HTTP-Protokoll aus dem WWW abrufbar sein. Sie unterscheiden sich in diesem Zusammenhang von den lokalen Schlüsseln, wie sie in einer Datenbank verwendet werden.
Zur Auszeichnung der Daten wurde oft XML angeführt. Die austauschbaren Wissensstrukturen sollten in RDF abgelegt werden. Um Daten zu beschreiben, werden Ontologien genutzt. Die Beschreibung grundlegender Datenstrukturen ist dabei durch das Vokabular RDF-Schema[44], kurz rdfs definiert. Das rdfs Vokabular gibt wichtige Zusammenhänge vor, wie beispielsweise die Beschreibung von Prädikaten, also der möglichen Zusammenhänge zwischen Fakten. Auch kann definiert werden, was diese Prädikate verbinden können. Das wird über die zulässigen Klassen der Subjekt- und Objektknoten geregelt. In der klassischen Datenbankwelt entspricht das der Tabellen- und Verknüpfungsstruktur.
Bei den Beschreibungsstandards für Ontologien hat sich die Web Ontology Language kurz OWL als Standard etabliert. OWL macht es möglich, Einschränkungen über das Klassensystem wie Kardinalitäten und sich ausschließende Klassen zu definieren. Sich gegenseitig ausschließende Klassifikationen sind wichtig, da in RDF Mehrfachklassifizierungen explizit erlaubt sind. In RDF können Entitäten beispielsweise gleichzeitig nach Funktion, ästhetischen Gesichtspunkten und Form klassifiziert werden. Des Weiteren können Prädikate, also die Verbindungen, als transitiv oder symmetrisch gekennzeichnet werden. Abschnitt 2.6.2 beschäftigt sich auch mit dem Thema. Alle diese Eigenschaften sind in relationalen Datenbanken und XML-Daten nicht ohne die implizite Erweiterung der Struktur möglich.
Die Abfragesprache für RDF ist SPARQL[112]. SPARQL unterstützt durch die Struktur von RDF einige spezielle Abfragefeatures, die sich stark von denen relationaler Datenbanken unterscheiden. Dabei ist jedoch zu bemerken, dass es im Standard nur begrenzte Abfragemöglichkeiten über Klassenhierarchien gibt. Hier gibt es diverse Erweiterungen über Reasoner, die jedoch nicht von allen SPARQL-Endpoints unterstützt, werden.[233]
Bis zu diesem Punkt kann der Semantic Web-Stack als ein Beschreibungs- und Modellierungsansatz gesehen werden. Die Regelsprachen, die hier auftauchen, kommen aus der Tradition der künstlichen Intelligenz. Hier gibt und gab es einige Standards, die mehr oder weniger mächtige Regeln definieren und verarbeiten konnten. An dieser Stelle fasert der Technologiestack auseinander. Es gibt bei den Regelsprachen einige Variationen von Sprachen und diese sind verschieden ausdrucksstark und komplex. Das hat auch mit der zugrunde liegenden Ontologie und ihrer Ausdrucksstärke zu tun.
Die letze konkret implementierte Schicht des Semantic Webstacks ist RIF, das Rule Interchange Format, welches als eine Art erweiterbares Grundgerüst für verschiedene Regeln gilt. Dabei wurden die Regeln in bestimmte Dialekte verschiedener Ausdrucksstärken zerlegt. [142]

4.1.2 Die Grenzen der Utopie

Die Grenzen der Utopie des Semantic Web sind an den Grenzen von Wissen und der Logik zu sehen. Dabei ist es wichtig, die Fallstricke der Logik und Abbildung zu kennen. Es ist im Falle eines Netzes von Information nicht einfach, Fragen zu stellen und Regeln zu formulieren, denn ein Netz aus Informationen muss nicht so aussehen wie die Regeln, die sich der Mensch für das Netz ausgedacht hat.

Der unvollendete Semantic Web-Stack

Der Semantic Webstack ist nicht bis an sein erdachtes Ende implementiert. Die unklaren Elemente lassen sich durch das reine Vergleichen der Stack-Visualisierungen erkennen. Hier gibt es die Elemente (Unified) Logic, Proof und Trust.
Die Schichten Logic und Proof dringen dabei weiter in die Domäne der künstlichen Intelligenz vor.
Eine weitere Schicht sollte das Vertrauen und die Überprüfung der Daten regeln. Nicht alle Datenquellen bieten die gleiche Datenqualität. Auch könnten Quellen manipulierte Informationen enthalten. Die Gefahr besteht darin, dass schlechtes oder fehlerhaftes Wissen bereitgestellt wird. Hier ließen sich im Blick auf Page-Rank-Manipulationsversuche auch ganz konkrete Gefahren anführen. Eine möglichst große, offene Welt zieht dabei Manipulation an. Daten könnten bereitgestellt werden, die logische Schlüsse in eine Richtung lenken. Gerade wenn sich eine geschäftliche Nutzung wie beim Zahnarzttermin ergibt, steigt das Risiko, dass Wissen zum eigenen Vorteil manipuliert wird.
Der Trust-Layer ist bei einem Informationsnetz von enormer, unüberschaubarer Größe sehr sinnvoll. In einem überschaubaren Rahmen ist dieser Layer aber auch noch gut händisch verwaltbar, beispielsweise durch eine White- oder Blacklist von Quellen.

Nutzbarkeit und Schlankheit

Die Vision des Semantic Web kann sich nur an der Realität messen und hier gibt es aus der Welt der Informationstechnologien gute Beispiele, die vor Augen führen, wie Technologie und deren Benutzung von der Nutzbarkeit und der tatsächlichen Nutzung abhängen.
Am Beispiel von HTML lässt sich zeigen, dass die simple Massenkompatibilität eines Standards über dessen Erfolg entscheidet. Jedem ist es möglich mithilfe kostenlos zugänglicher Quellen und ohne unfangreiches Vorwissen HTML zu schreiben. Fast alle Textverarbeitungswerkzeuge haben eine HTML-Export Funktion. Das Bereitstellen von HTML-Inhalten über einen eigenen Server oder dem Platz bei einem Webhoster ist sehr einfach. Das hat es vielen Menschen ermöglicht, Informationen mittels HTML bereitzustellen, auch wenn sie über wenige technische Mittel verfügen.
Das hat die Verbreitung als Standard vorangetrieben, auch wenn es jahrelang Kompatibilitätsprobleme bei der Interpretation von HTML und anderen Technologien aus dessen Umfeld aufseiten einzelner Browser gab.
Statisches HTML wird nicht ausgetauscht, um kollaborativ Texte zu schreiben. Hier gibt es Systeme wie Wikis, die für solche Aufgaben den Hypertext verwalten. Dabei wird bei

Wikis noch einmal eine Komplexitätsreduktion vorgenommen und eine mehr oder weniger angepasste Wiki-Syntax verwendet. Diese ist dann wieder ein Simpler funktionaler Dialekt und wird in HTML abgebildet.
Der Standard SGML[4], welcher grob als eine komplexe Art von XML und HTML beschrieben werden kann, hat sich nicht durchgesetzt. SGML wurde durch den einfacheren Standard XML ersetzt. Die niedrige Ausdrucksstärke war anscheinend kein Hindernis für dessen Erfolg, auch wenn anfangs geplant war, XML und SGML kompatibel zu machen.[58] Anscheinend waren die Features von XML für die meisten Nutzer ausreichend. Außerdem war eine Anpassung an die Gegebenheiten eines neuen Umfelds, dem Internet und einer größeren potenziellen Nutzerschaft durch die Ähnlichkeit zu HTML ausschlaggebend für den Erfolg.[84]
Die hier beschriebenen Erfolgsmodelle stehen im Widerspruch zur Ausrichtung auf das abstrakte Ziel, künstliche Intelligenz, zu schaffen wie es das Semantic Web verfolgt hat. Um eine allgemein nutzbare Technologie zu schaffen, muss sie für menschliche Nutzer mit einer breiten Basis handhabbar sein.

Komplexe Beschreibung

Das Semantic Web ist eine Initiative, die sehr von der Wissenschaft und wissenschaftlichen Institutionen vorangetrieben wurde. Das hatte auch Konsequenzen in der Definition von Ontologien. Diese Ontologien wurden sehr komplex.[5] Diese Komplexität macht laut Shadtbolt et al.[224] die Ontologien nicht wirklich verwendbar. Es erfordert hohen Aufwand, etwas in ihnen abzubilden.
Konkret kann das Problem so beschrieben werden: Komplexe Klassifikationen verlangen viel Aufwand beim richtigen Klassifizieren. Es müssen komplexe Systeme verstanden und soweit durchdrungen werden, damit diese zum Auszeichnen von Daten verwendet werden können.
Um vergleichbare Daten zu erschaffen, muss die Datenmodellierung von der anderen Seite ähnlich angegangen werden. Des Weiteren müssen die Daten aus zwei verschiedenen Quellen Attribute aufweisen, die vergleichbar und aufeinander abbildbar sind.
Das Gleiche gilt nochmals für denjenigen, der am Ende die Daten zur Erstellung von Regeln oder zur Programmierung von Anwendungen verwenden will.
Dass am Ende wirklich etwas Vergleichbares herauskommt, ist somit schwer zu realisieren, da Daten zu bestimmten Zwecken, in bestimmten Sprachen und mit ausgewählten Attributen gesammelt werden. Diese Daten passen am Ende doch nicht zusammen oder entsprechen dann nicht den Erwartungen der Nutzer. In diesem Fall wird eine Ontologie zudem schlecht nutzbar, da die Analyse der Daten in der Abbildungspraxis schwer wird.
Die Partizipation an der Gestaltung von Ontologien ist dabei auch ein Politikum und kann nicht einfach auf die Güte oder Vollständigkeit der Abbildung beschränkt werden. Es geht hier um die Definition eines Byte gewordenen “Common Sense”. Es kann daher

[4]Standard Generalized Markup Language

[5]Bei Shadbolt auch “Deep Ontologies” genannt[224]

einzelnen Parteien, politisch oder wirtschaftlich, opportun sein, die eigene Weltsicht als allgemeingültig zu definieren. Hier würde eine Ontologie im Sinne eines Standards verwendet.
Wenige könnten eine Ontologie ihren Nöten anpassen, sodass die Transaktionskosten, die eigenen Daten umzuwandeln, geringer werden und große Teile ihres Datenbestands ohne Verluste abbilden können, während andere zuarbeiten müssen. In diesem Fall würde eine Ontologie zum informationellen Machtinstrument.

Offene Welt

Bei Datenbanken wird oft von einer geschlossenen Welt[6] gesprochen. In einer geschlossenen Welt ist das Wissen vollständig. Alles, was in der Datenbank nicht existiert, gibt es nicht. Das Semantic Web und Linked Data gehen davon aus, dass wenn etwas nicht erfasst ist, es entweder nicht stimmt oder nicht bekannt ist. Gerade im Hinblick auf Regelsysteme ist diese Annahme das wichtig. Eine Falsifikation ist in der offenen Welt schwer durchzuführen, da keine Möglichkeit des Ablegens von negativem Wissen in RDF besteht. Negatives Wissen würde eine immense Größe erreichen und ist daher nicht vorgesehen.[134, 172]
Allein die Information, wessen Kind jemand nicht ist, würde immens komplex zu beantworten sein. Die Frage ist dabei auch, ob allgemein abgebildet werden muss ob Eltern bekannt sind. Einen pragmatischen Ansatz, dieses Problem zu lösen, bildet die Einschränkung der Kardinalität von Beziehungen wie in relationalen Datenbanken. Dies ist in OWL möglich. Eine Elternbeziehung wäre damit auf eine Mutter und einen Vater beschränkt, wobei von biologischen Eltern ausgegangen wird. Das mag eine Einschränkung sein, die einleuchtet. Aber diese Abgrenzung könnte dabei auch ein Problem aufzeigen, denn bei nichtbiologischen Eltern oder unklaren Verhältnissen dieser Art kann es zu Informationsmodellierungsproblemen kommen. Natürlich lässt sich diese Fragestellung weiterhin durch den Fortschritt der Reproduktionsmedizin beliebig komplex gestalten. Kurz: Unsicherheit bei der Modellierung kann selbst in diesem, eigentlich recht eindeutig wirkenden Beispiel auftreten.
Die Frage nach der Vollständigkeit ist auch bei transitiven Fragestellungen problematisch. So kann es bei fehlendem automatischen Schließen nützlich sein, wenn transitive Beziehungen explizit abgebildet werden. Um im Beispiel zu bleiben: Jede Großeltern-, Urgroßelternbeziehung u.s.w. müsste abgebildet werden. Daraus würden sehr viele Verbindungen entstehen. Diese Beziehungen könnten aber auch immer wieder durch Inferenz hergeleitet werden. Hier stellt sich die Frage der Verhältnismäßigkeit der expliziten Speicherung dieses eigentlich angestrebten “neuen” Wissens.

Komplexe Welt, komplexe Regeln

Die Beschreibung durch komplexe Ontologien wirkt sich auch auf Regelsysteme aus.
Die Komplexität von Regeln und des Datenmodells spielt eine zentrale Rolle bei der Berechnungszeit von Problemen. So sind nicht alle Schlüsse, die in den verschiedenen OWL-

[6] Engl.: closed-world

Standards beschrieben werden, immer in absehbarer Zeit lösbar. Als weiteres Problem kommt dabei die Größe der verfügbaren Daten hinzu.
Wenn eine Ontologie sehr komplex ist, wird es natürlich auch schwerer, Wissen aus ihr herzuleiten, wie es das Ziel von Regelsystemen ist. Dabei können Regeln anhand einer Ontologie erstellt werden. Dies birgt die Gefahr, dass eine Menge von Regeln formuliert wird, die am Ende nicht angewendet werden, da keine Daten existieren, auf die diese Regeln passen. An dieser Stelle trifft eine ideale Welt auf eine tatsächliche Welt.

4.1.3 Linked Resignation

Die rationale Betrachtung des Semantic Web-Ideals lässt sich unter Linked Data zusammenfassen. Linked Data bedient sich der Technologien des implementierten Semantic Web. Dabei geht es in erster Linie um die Verbindung von Datenbeständen und das Zugänglichmachen von Informationen in maschinenlesbarer Form.
Im Gegensatz zum Semantic Web ist Linked Data pragmatischer. Die Linked Data-Prinzipien sind dabei recht einfach zu verstehen und umzusetzen.

1. URIs werden als Namen für Dinge verwendet
2. URIs können über das HTTP-Protokoll aufgelöst werden
3. Hinter den URIs befinden sich nützliche Informationen, die den Linked Data-Standards folgen
4. Diese Informationen enthalten mehr URIs zu anderen Dingen

[128, 26, Kapitel 2] Diese einfachen Regeln der Linked Data-Bewegung klingen gerade in der ersten und der letzten Regel wie ein algorithmisch rekursives Mantra. Dinge verknüpfen sich mit Dingen. Dieses Prinzip bildet schon aus sich heraus ein Netz, ohne dass dabei ein Netz intendiert sein muss.
Linked Data geht weg von den komplexen Strukturen, wie sie in der Vorstellung des Semantic Web formuliert wurden. Starke Formalisierungen und die Komplexität der Ontologien machen das Bereitstellen von Daten in dieser "Sprache" aufwendig. Das schreckt bei der Verwendung ab und es wird schwerer, eine kritische Masse an Daten zu finden, welche die gleiche "Sprache" sprechen.
Dabei ist klar, dass komplexe Ontologien auch schwerer verständlich und aufwendiger in der Verarbeitung sind als eine simple Ontologie.
Linked Data ist eher eine pragmatische Verbindung von Wissen. Auf diese Weise sollen Informationsressourcen referenzierbar und vernetzbar werden. Der Vorteil, der daraus entsteht, ist ein ständig wachsendes Netz aus Informationen. Die Informationen sind relational aufeinander bezogen und es gibt Grunddaten, auf die verschiedene Datensätze referenzieren können.
Statt starre, komplexe Ontologien werden bei Linked Data eher einfache Ontologien oder Begriffsstrukturen aus Bottum-Up-Verfahren, wie Folksonomies als Austauschsprachen gesehen.[224] Diese sind dabei jedoch nicht einfach zu verstehen und die Analyse und Einordnung von Klassen kann unscharf sein.

Aus graphentheoretischer Sicht können Ontologien mindestens als bipartite Graphen gesehen werden und Folksonomies sogar als tripartit, da eine menschliche Komponente hinzukommt, die bei einer Ontologie eher hinter dem Anspruch der generellen Gültigkeit versteckt wird.[169]
Linked Data ist nicht nur im Geiste ein Kind des Semantic Web. Die Technologien des Semantic Web wurden weitestgehend übernommen. Daten aus dem Bereich des Semantic Web können mit Linked Data vernetzt werden.

4.1.4 Die Welt des sehr großen Wissens

Durch die linked Open Data-Bewegung und das Semantic Web werden eine Vielzahl von Informationsquellen bereitgestellt, dabei wird ein starker Fokus auf das Vernetzen von Wissen gelegt. Dies stellt Möglichkeiten bereit, was die Zweitnutzung von Daten angeht.
Auch hier spielt natürlich die Idee herein, dass "neues"[7] Wissen erschlossen werden kann. So können natürlich auch die Daten, in Form und Fokus, etwas über die Produzenten erzählen. Aber selbst ein Korpus wie Wikipedia stellt durch seine schiere Größe interessantes Wissen zur Verfügung. Die Menge der potenziell verfügbaren Daten kann durch die Vernetzung sehr groß werden. Ein Problem ist, interessante Daten überhaupt zu finden. In diesen Datenmengen sind Emergenzen nicht wirklich auf den ersten Blick ersichtlich.
Das Aufheben der Grenzen von Daten hat dabei eine weitere Konsequen: Datenstrukturen müssen nicht alles selbst beschreiben. Eigene Daten sind für die Vergleichbarkeit zwischen Datenquellen hinderlich. Es müsste bei einem globalen Abgleichen aller Datenbanken ein Abgleich aller Datenbanken mit allen anderen Datenbanken erfolgen. Wenn es aber eine zentrale Datenbank gibt, auf die sich alle berufen, dann könnte die Klassifikation genau über diese Datenbank geschehen. Alternativ könnten alle Datenbanken ihren eigenen Bestand mit dieser zentralen Datenbank abgleichen.

Linked Data Cloud

Die linked Open Data Cloud ist eine Wissenskomponente, die um DBPedia herum gebildet wurde. Die Initiative versucht, offene Informationsquellen miteinander zu verbinden. Als Mittler bzw. Hub dient dabei DBPedia.
Beim Auffinden und Vergleichen von Daten kommt es oft zum Problem der Identitätsbestimmung. Die eindeutige Identität eines Datensatzes oder einer Entität mit geringer Fehlerwahrscheinlichkeit zu ermitteln, kann sehr aufwendig sein.
Das ist in der Wikipedia insofern gelöst, als dass die Unterscheidung der Sachverhalte durch menschlichen Eingriff erfolgt. Ein Wikipediaartikel beschreibt einen Sachverhalt. Durch die Offenheit von Wikipedia entsteht so eine große Menge von unterschiedlichsten Inhalten. Das ist ein Grund, warum DBPedia als Zentrum der Linked Data-Cloud gewählt wurde.[11]
Die Verbindungen, die hier auf Wikipedia verweisen, werden oft durch Dateien repräsentiert

[7] Es lässt sich streiten, ab wann etwas Wissen ist und wann es neu ist. Der Begriff "neu" ist natürlich problematisch, da nicht jede Abbildung und jedes Zwischenergebnis relevantes neues Wissen darstellt.

die Listen von Referenzen bereitstellen. Es wird abgebildet welche URIs der offenen Datenbestände mit denen der DBPedia übereinstimmen. Technisch sind es Dateien mit RDF-Tripeln, die über das "OWL:SameAs" Prädikat eine Verbindung herstellen. Dabei handelt es sich meistens um einfache eindeutige eins zu eins Zuordnungen der URIs aus der Datenquellen zu URIs aus der DBPedia.
Im Hinblick auf die Popularität von Wikipedia ist der Schluss, DBPedia als zentralen Hub des Netzwerks von Linked Data zu verwenden, schlüssig. Es ist bei einem solchen System wichtig, dass es auf ein gemeinsames Vokabular zurückgreift. Durch seine allgemeine Ausrichtung sind die Daten in DBpedia bzw. Wikipedia dafür sehr gut geeignet.
Die Konstruktion von DBpedia als Datensatz wird in Abschnitt 4.2.2 genauer beschrieben.

Vollständigkeit

Linked Data und das Semantic Web gehen von einer offenen Welt, einer Welt von verschiedensten Datenquellen aus. Wie kann von der Vollständigkeit von Daten gesprochen werden?
Dies ist ein wichtiges Problem, wenn Daten unter dem Aspekt der stochastischen Validierung gesehen werden. Dabei ist zu beachten, dass manche Daten schneller und einfacher den Weg in die Linked Data Cloud finden werden als andere. Somit ist davon auszugehen, dass die Verteilung von Wissen im Netz der Informationen nicht zufällig und gleichmäßig verteilt ist.
Erkannt wurde das Problem in der Community der linked Open Government Data. Hier ist eine Forderung, dass die zu veröffentlichten Daten möglichst vollständig sein müssen.
Auch gibt es eine Verteilung von Wissen im Sinne des Power-Law. Über populäre Dinge existieren wahrscheinlich mehr Fakten als über weniger populäre Sachverhalte. In Fällen wie der Wikipedia kann ein Sachverhalt auch von einer kleinen Gruppe im Detail dargestellt werden, während andere Bereiche nicht gut dokumentiert sind.
Das ist eine Konsequenz der kollaborativen Struktur von Wikipedia. Da Allgemeinwissen schwer zu definieren ist und auch über die Zeit in der Definition variieren kann, ist eine Vollständigkeit hier noch schwerer zu erreichen. Wer sollte Allgemeinwissen oder enzyklopädisches Wissen definieren? Das ist ein Problem, das sich in der Wikipedia beispielsweise in der Diskussion um relevante Artikel zeigt.
Diese Faktoren machen es schwer, Wissen aus vernetzten Daten mit RDF zu generieren bzw. Objektive und vollständige Daten aus Linked Data Quellen zu ziehen.

4.1.5 Zusammenfassung

Die Auszeichnung und Nutzung von Linked Data hat einige Probleme zur Verwendung der Daten für die klassische Social Network Analysis (SNA). Um einen Aspekt zu analysieren, muss erkannt werden, was konkret bereitsteht. Das liegt jedoch in gewisser Weise im Verborgenen. Für die Auswertung von Daten stellt sich die Frage, wo interessante Klumpen von Wissen auftreten. Noch ist eine wirkliche Abfrage, wie durch den Menschen, nötig, um zu sehen, wo interessantes Wissen vorliegt. Wichtig ist auch, was die hinterlegten Informa-

tionen im Genauen bedeuten. Das kann im Text der Datenquelle oder den Originaldaten erschlossen werden.

4.2 Datenherkunft und Datenbereitstellung

Die Daten aus dem Bereich des Semantic Web und Linked Data sind, abgesehen von ihrer intendierten Verwendung, aus mehreren Gründen im Rahmen dieser Arbeit interessant. Für den Nutzer wird eine mehr oder weniger einheitliche Datenstruktur bereitstellt. Erste Normalisierungen an eine offene Welt sind vorbereitet. Die Datenstrukturen sind an sich an einer Netzstruktur orientiert. In der Datenerstellung sind auch interessante Ansätze zu sehen, wie Daten in ein Netz umgewandelt werden können. Auf der anderen Seite haben gerade verteilte Daten, die zwischen verschiedenen Datenprovidern verlinken, spezielle Probleme.
Ein Link ist dabei nur so gut wie das Ende, an dem man ankommt, wenn man ihm folgt. Wenn ein Provider dauerhaft ausfällt, dann hängt das Ende eines Links in der Luft, diese Links werden auch tote Links genannt.
Der Ausfall von Links ist dabei besonders kritisch, wenn Datensätze, wie beschrieben, keine Teildatenbestände oder Referenzen auf andere Datenquellen erstellen, um ihre Daten zu vervollständigen.

4.2.1 Datenerstellung

Die Daten im Semantic Web und in Linked Data sind meistens aus anderen Standards erstellt und werden über normale Datenbanken verwaltet.
Die Daten werden regelmäßig aus den Ursprungsdatenformaten anhand eines Mappings in RDF exportiert. Formate wie Tabellen, Artikel, Geodaten und strukturierter Text können in den Semantik Web-Standards abgebildet und vernetzt werden. So gibt es beispielsweise für die Gesetze, wie sie in Abschnitt 3.2.2 untersucht wurden, Vorschläge für die Umwandlung in RDF.[222]

Relationale Datenbanken

Eine Möglichkeit Linked Data zu erzeugen, ist die Überführung von relationalen Datenbanken auf Linked Data-Standards.
Bei diesem Unterfangen geben sowohl die Ontologien oder Begriffssysteme, die zur Auszeichnung verwendet sollen, als auch die Komplexität des Quelldatensatzes die Komplexität vor. Wie schon erwähnt lassen sich, je nach Komplexität, die Quelldaten nicht komplett auf die Ontologien abbilden. Das wird gerade bei Linked Data in Kauf genommen, da hier auf schlanke Ontologien gesetzt wird. Bei komplexen, vollständigen Abbildungen ist die korrekte Modellierung sehr aufwendig.
Eine Technologie, die dabei hilft, Datenbanken als Linked Data verfügbar zu machen, ist D2R. Sie ermöglicht es, geschlossene Datenbestände in Linked Data im RDF-Schema zu publizieren.[32] Die Schnittstelle bildet ein Mapping in einer XML-basierten Befehlssprache. Diese ist so angelegt, dass sie SQL-Befehle in eine Menge von RDF-Aussagen

überführt. Dabei werden die URIs dynamisch aus den Schlüsseln der Daten erstellt.[31] Entitäten lassen sich an dieser Stelle einfach aus der Datenbank herausgeben. Solange eine Typisierung der Daten machbar und durch eine Ontologie in der Klassenstruktur abbildbar ist, kann der relationale Teil einer Datenbank einfach übertragen werden. Dabei ist jedoch auch wichtig, dass die Prädikate eindeutig aus den Relationen hervorgehen, ansonsten ist die eindeutige Klassifikation einer Relation problematisch.

Hypertext, XML

Eine andere Art der Datenerstellung ist die Umwandlung von Hypertext oder XML. Dokumente können auf ihre Links reduziert werden. Dabei würde eine sehr basale Relation entstehen. Andererseits können auch strukturelle Eigenschaften von Dokumenten nachgebildet werden.[222]
Es kann aber auch versucht werden, semantisch typisierte Daten aus diesen Dokumenten zu entnehmen. Das kann beispielsweise bei strikt gegliederten Dokumenten funktionieren. Zum Umformen können dabei Technologien wie XSLT und Xpath helfen.
Ein Standard, um HTML direkt in RDF zu verwandeln, ist RDFa[4], welches eine Annotationserweiterung für HTML ist. Dabei ist das Subjekt im RDF Triple immer das Dokument. Das Ziel des Links ist das Objekt und RDFa ermöglicht es, die Beziehung zwischen Ausgangs- und Ziel-Dokument im HTML-Tag zu typisieren.

4.2.2 Beispiel DBpedia

Das DBpedia-Projekt[33] ist eines der bekanntesten Linked Data-Projekte und wie beschrieben der zentrale Hub der Linked Data Cloud. Wikipedia hat aufgrund der Wikistruktur einige Eigenschaften, die viele verschiedene Umwandlungsformen von Linked Data zulassen. Dabei bietet die Mischung aus einer relationalen Datenbank sowie speziell typisiertem Hypertext viele Möglichkeiten der Extraktion von vernetzten Informationen.
Eine Datenfacette sind dabei alle Links von Wikiseiten, wie sie später auch verwendet werden. Dies sind ganz normale Links im Sinne einer URL und beinhalten keine semantischen Daten. Die semantische Codierung der Daten fängt an, wenn der Kontext von Links einbezogen wird. So werden Page-Links und Links zu anderen Websites unterschieden. Ein Artikel kann dabei einfach einen anderen Wikipediaartikel oder eine Seite außerhalb von Wikipedia verlinken. Die Links können als Linked Data bezeichnet werden, da sie DBpedia Artikel untereinander und mit anderen Ressourcen verbinden. Die Verbindungen, die dabei entstehen, sind nicht stark und restriktiv definiert.
Das reduziert ihren spezifischen Informationsgehalt, senkt dabei jedoch auch die Komplexität der Daten. Die Komplexität einer Beziehung ist meist von der Ontologie abhängig, die zur Auszeichnung verwendet wird.
Eine wichtige Quelle für semantische Informationen in DBpedia sind die Info-Boxen.[33] Die Struktur der Info-Boxen macht es einfach, Informationen zu extrahieren und daher RDF-Tripel zu bestimmen. Es gibt von der Community je nach Typ des Wikiartikels festgelegte, strukturierte Vorlagen für Info-Boxen. In diesen Vorlagen sind verschiedene Arten von Feldern definiert. Diese stellen eine Bottom-up-Ontologie oder schwache Begriffsstruk-

tur dar.
Dabei können die Werte in diesen Info-Boxen durch reinen Text, Werte wie Koordinaten und Zahlen oder andere Wikipediaartikel durch einen internen Link referenziert sein. Die Daten sind einer Key-Value-Beziehung, in denen die Values durch Zeichenketten (Freitext) abgebildet werden, sehr ähnlich. Diese Zeichenketten können auch komplexere Daten enthalten. Ein typisiertes RDF-Tripel lässt sich dabei aus der Kombination der Identität des Artikels und einem Eintrag in einer Infobox darstellen. Der Artikel dient als Subjekt, der Key der Tabelle als typisiertes Prädikat. Das Objekt kann ein anderer URI oder ein Freitext sein, je nachdem, ob ein anderer Artikel in der Infobox verlinkt ist oder nicht. Das Vorgehen erinnert an die Auszeichnung von HTML-Links mit RDFa. Ansonsten wird das Objekt durch einen String oder einem daraus hergeleiteten Wert repräsentiert.
Personen zugeordnet sind beispielsweise Sterbe-, Geburts- und Wirkungsorte sowie Datumsangaben. Als Verfeinerung kommen bei Monarchen beispielsweise Nachfolger oder Regierungsgebiete als Eigenschaften hinzu. Hier lassen sich die Geburts- und Sterbeorte als Entitätsbeziehungen sehen, wenn die Orte entsprechend verlinkt sind. Daraus lassen sich komplexe Visualisierungen erstellen.[220]
Diese Informationen sind anscheinend recht zuverlässig dokumentiert und oft unumstritten, daher werden sie oft als Links angegeben. Ein Problem stellt hier die Granularität der Beziehung dar. Es kann sich bei der Angabe eines Links um ein Land, eine Stadt, eine Region oder ein Krankenhaus handeln. Auch die Nennungen mehrerer Orte auf verschiedenen Genauigkeitsstufen sind möglich, aber nicht verpflichtend.
Eine andere Möglichkeit der Wertausprägung sind Koordinaten, wie sie bei Orten selbst und Bauwerken sehr sinnvoll sind. Dabei ist zu beachten, dass auch hier ein Zwiespalt zwischen Entität und Wert vorliegt, da ein zeitliches Datum, also beispielsweise ein Jahr, auch als Entität aufgefasst werden kann und eine Koordinate einem benennbaren Ort zugewiesen werden kann. Die Daten können verschieden genau in der Wikipedia abgelegt sein: Als Jahr in Form einer Zeichenkette, als Jahr in Form eines Datums und als Datum in Form eines genau bestimmten Tages.
In komplexeren Fällen können aber auch reine Zeichenketten an diesen Positionen stehen. Das ist in einer offenen, eigentlich menschenlesbaren Quelle wie Wikipedia auch vollkommen legitim. Für die Abbildung in maschinen lesbaren Daten hat das jedoch den Nachteil, dass für die Verarbeitung und den Vergleich von Zeichenketten mehr Aufwand betrieben werden muss und zwangsweise neue Fehler entstehen. Auch ist nicht klar, inwieweit eine Information konkretisiert werden kann, wenn selbst eine Primärquelle in dieser Hinsicht unscharfe Informationen liefern würde.
Die Semantic Web-Daten stellt das DBPedia-Projekt durch einen SPARQL-Endpoint zur Verfügung, aber auch als statisch, versioniertes Datenpaket. Der SPARQL-Endpoint ist eine Schnittstelle, die Anfragen an die Datenbank beantwortet. SPARQL-Anfragen sind dabei eine Möglichkeit, interessante Muster aus dem riesigen Korpus zu extrahieren. Der SPARQL-Endpoint des DBPedia-Projekts stellt dabei jedoch nicht die einfachen Wikipedia Page Links zur Verfügung, sondern die typisierten Links aus den Info-Boxen.
Im Kontext von Wikis gibt es weitere Möglichkeiten der Datenumwandlung, wie das se-

mantic Wiki Projekt.[148, 216] Es ist eine Erweiterung für das von Wikipedia verwendete Media Wiki. Es ermöglicht, direkt bei der Erstellung der Wiki-Inhalte semantische Datenstrukturen in den Info-Boxen und den Links zu hinterlegen.

4.2.3 Vernetzen

Das Vernetzen des in der Ausgangsform verteilten Wissens ist eine der grundlegendsten Aufgaben im Bereich von Linked Data. Hierzu gehört beispielsweise das Verbinden zweier Datensätze aus verschiedenen Datenbasen, welche dasselbe beschreiben. An dieser Stelle ist es auch wichtig zu unterscheiden, ob es sich bei den verlinkten Datensätzen um dasselbe, also ein Ding der gleichen Kategorie, oder nur um das Gleiche, das exakt gleiche Objekt einer Kategorie, handelt. Diese Aufgabe an sich ist nicht immer ganz einfach, selbst wenn die Datenstrukturen der Datenbasen sehr ähnlich sind.
Da hier Wikipedia bzw. DBpedia als Beispiel dient, muss gesagt werden, dass Wikipedia eine stark genutzte Online-Quelle mit vielen Nutzern ist. Auch wenn der Korpus groß ist, können viele Verlinkungen durch die große Nutzerzahl manuell angelegt und gesichtet werden. In anderen Projekten ist diese Arbeitskraft nicht vorhanden. Daher werden Linked Data Datensätze zwischen verschiedenen Datenquellen meistens nicht per Hand angelegt. Viele Datensätze sind zu groß und die manuelle Verlinkung wäre händisch nicht zu finanzieren. Das automatische Vernetzen von Datensätzen war daher schon in der Frühzeit der computergestützten Informationsverarbeitung ein wichtiges Thema.[179]
Einen Überblick über das Feld der Vernetzungstechniken verschiedener Datenbanken bietet Kummer[151] am Beispiel der Vernetzung der Arachne Datenbank und der Perseus Digital Library. Bei einem Abgleich von zwei Datenbasen müssen alle Elemente in der ersten Datenbasis mit den Elementen in der zweiten Datenbasis verglichen werden. Dies wird Matching genannt. Bei großen Datenmengen kann das selbst mit moderner Technik sehr viel Zeit in Anspruch nehmen. Für eine Verbesserung der Performance und zur Vermeidung von Fehlern können Datensätze in Untergruppen unterteilt werden, die prinzipiell vergleichbar sind. Diese Technik wird als “Blocking” bezeichnet.
Auch kann Freitext bei dieser Aufgabe helfen. Dieses Vorgehen wurde in Abschnitt 3.1.3 genauer beschrieben. Im Fall von DBpedia Spotlight [167] gibt es ein eigenes Framework für die Annotation von Entitäten aus dem DBpedia-Projekt. Die Tests zu Precision und Recall fallen bei DBpedia Spotlight eher schlecht aus.[144] Dabei ist zu beachten, dass alle Entitäten in DBPedia gesucht werden und andere NER Frameworks mit spezielleren Zielmengen getestet werden. Die Möglichkeit, Falsches zu erkennen, ist bei DBPedia Spotlight höher, da es mehr mögliche Kandidaten gibt, die erkannt werden können.

Wikipedia Language Links

Ein Beispiel für das Verlinken von verschiedenen Datenquellen zeigt sich im Verbinden von Artikeln zwischen den verschiedensprachigen Versionen der Wikipedia. Hier wird mir sogenannten Language-Links gearbeitet. Diese Language-Links verbinden die verschiedenen Sprachversionen von Wikipedia. Dabei kann es zu uneindeutigen, spezifischen und widersprüchlichen Zuordnungen kommen. So muss ein Thema im Deutschen nicht in der

gleichen Granularität vorliegen wie im Englischen, hätte ein Thema im Englischen einen ganzen Artikel, während in der deutschen Wikipedia nur der Absatz eines allgemeineren Artikels vorliegt.
Auch die Entstehung von Daten hat Einfluss auf das Verlinken eines Artikels. Handelt es sich bei einem neu angelegten Artikel um eine ergänzende Übersetzung eines Artikels aus einer anderen Sprachversion, dann ist ein Link einfach und eindeutig.
Werden zwei Artikel, die vorher unabhängig voneinander erstellt wurden, im Nachhinein verlinkt, kommt es wahrscheinlich eher zu Granularitätsproblemen oder sogar Bedeutungsproblemen. Des Weiteren muss ein Mensch, der einen Wikipedia Language-Link erstellt, den Sprachen an beiden Enden des Links mächtig sein. Bei einer Evaluation mehrerer Links wird das Problem noch größer, da die genauen Bedeutungen in eine Vielzahl von Sprachen auftreten können. Hier können Cluster-Analysen dabei helfen, Probleme aufzudecken.[244]
Zur Lösung der hier beschriebenen Vernetzungsprobleme wurden auch einige automatische Verfahren vorgeschlagen.[237, 187] Im Vorschlag von Sorg und Cimiano wird eine Heuristik vorgestellt, die Kandidaten aus bestehenden Links herzuleiten. Dabei wird angenommen und auch empirisch bestätigt, dass es ein Muster gibt, das die Kandidaten für Übersetzungen einschränken kann. Die Übersetzung eines Artikels in einer anderen Sprache findet sich dabei im Kontext eines Artikels, der durch einen Language-Link übersetzt werden kann.[237] Hier wird aus vorhandenem Wissen eine Heuristik für neues Wissen hergeleitet. Das Verwenden solcher Verfahren erhöht potenziell den Informationsbias der Daten. Dinge, über die viel bekannt ist, in Form von Links und Text, können weiter verbunden werden, Diese Phänomene spiegeln sich auch in der der Longtail-Verteilungen, wie sie in Abschnitt 2.5.6 beschrieben wurden wieder.
Andere Auswertungen der Language Links abstrahieren von den konkreten Artikeln und bewerten die Links nach den Sprachen, die sie verbinden. Aus diesen Informationen lässt sich ablesen, welche Sprachkombinationen von Menschen genutzt werden.[213]

Abgleichen von Daten am Beispiel von Pelagios

Das Pelagios-Projekt beschäftigte sich mit dem Verknüpfen von Daten auf Ortsebene[231, 137]. Ziel des Projektes in seiner zweiten Phase war es, verschiedene Datenquellen anhand ihrer Ortsinformationen miteinander zu vernetzen.
Dabei wurden verschiedene Projekte aus den Altertumswissenschaften einem Gazetteer, Pleiades[199], zugeordnet. Hierzu gehörte auch die Arachne-Datenbank, Wie schon bei der DBPedia hat dieses Hub oder Stern-Vorgehen den Vorteil, dass auf ein einziges System verwiesen wird. So können am Ende alle Systeme über dieses gemeinsame Referenzsystem, den Pleiades-Gazetteer, ihre Daten anhand von Orten vergleichen. Der Vorteil liegt in der sternförmigen Verlinkung. Wie bei einer Projektion eines bipartiten Netzes können theoretisch die einzelnen verlinkenden Systeme direkt untereinander verbunden und vergleichbar gemacht werden. Dabei muss für jede angebundene Datenquelle nur ein Datenabgleich gemacht werden. Würden alle Datenquellen untereinander verlinken, würde jede Datenquelle mit jeder anderen abgeglichen werden müssen. Statt einer linearen Menge von Verlinkungs-

initiativen müsste eine quadratische Menge von Verlinkungen angegangen werden.
Bei dem Zuordnen von Orten kommt es dabei zu speziellen Problemen.
Die Definition vieler Orte hängt dabei vom Zeitpunkt ab. So kann ein Land zu verschiedenen Zeitpunkten unterschiedlich groß sein. Länder können zerbrechen oder Territorien aus ihrem Hoheitsgebiet verlieren. Ein gutes Beispiel dafür ist der schwammige Begriff Deutschland. Im historischen Kontext könnte Deutschland ein Kulturraum sein. Es kann als Deutsches Reich und als Deutschland in den Grenzen zu einem bestimmten Zeitpunkt verstanden werden. Hinzu kommt die deutsche Teilung, in der es zwei politische Einheiten mit dem Namen Deutschland gab. Die Deutsche Demokratische Republik und die Bundesrepublik Deutschland. Es gibt also eine Vielzahl von Definitionen, die immer andere Landmassen umfassen. Diese Aufspaltung und Einteilung macht natürlich auch Probleme bei der Einstellung von Linked Data und Semantic Web Daten. Eine einfache Einteilung in Deutschland würde dabei eine Verlinkung zu den Nachbarstaaten schwierig machen. Für die DDR wäre Polen ein Nachbarstaat, nicht aber Frankreich. Umgekehrt wäre es für die BRD der Fall. Diese Ungenauigkeit könnte für die Inferenz über komplexe Regelsystem ein Problem darstellen.
Das direkte Zuordnen von Orten könnte auch über die Koordinaten erfolgen. Das birgt aber auch Probleme. Punktbeschreibungen, die nicht aus der gleichen Primärquelle oder Messung stammen, liegen hier nur in den seltensten Fällen wirklich aufeinander. Es hängt dabei von der Wahl des Mittelpunktes und der Genauigkeit einer Koordinate, ab, inwiefern solche Daten vergleichbar sind. Für die Zuordnung ist es auch wichtig, inwieweit es mehrere Kandidaten für eine Zuordnung in den Datensätzen geben kann. Nicht immer ist eine eineindeutige Zuordnung möglich oder sinnvoll. Außerdem hat ein Ort immer eine gewisse definitorische Ausdehnung, die implizit oder explizit[8] ausgezeichnet sein kann.
Eine andere Beschreibungsart bietet eine Fläche in Form eines Polygons. Dies verschiebt das Problem der Identitätsfindung nur auf eine andere Stufe. Hier müssten die überlappenden Flächen gefunden werden. Diese würden dann aber auch voneinander abweichen, wie die einzelnen Punkte es schon tun.
Im Ganzen ist es unumgänglich, beim Zuordnen verschiedener geografischer Datensätze eine gewisse Art des Fuzzy-Matchings bzw. einen Toleranzrahmen, zu nutzen. Wichtig ist auch die semantische Eindeutigkeit einer Verbindung. Beziehungen über ein "SeeAlso"[9] können recht unproblematisch erstellt werden. Verbindungen mit Prädikaten wie "falls_within"[10] , "overlaps_with"[11] oder "sameAs"[12] sind viel schwerer zu bestimmen. Alle sich daraus ergebenden Mappings werden entweder mehr Fehler enthalten (niedriger Precision Wert) oder sehr dünn werden (niedriger Recall Wert) und damit ihren allgemein integrativen Charakter verlieren.[147, 146] Bei einer Zuordnung über Namen können ver-

[8] Beispielsweise die Angabe eines Radius zu einem Punkt.

[9] `https://www.w3.org/TR/rdf-schema/#ch_seealso`

[10] Diese Prädikate kommen aus dem CIDOC-CRM[76] Standard und werden hier beispielhaft verwendet.`http://erlangen-crm.org/current/P89_falls_within`

[11] `http://erlangen-crm.org/current/P121_overlaps_with`

[12] `http://www.w3.org/TR/2004/REC-owl-semantics-20040210/#owl_sameAs`

schiedene Schreibweisen und verschiedene Sprachen das Matching erschweren. Andererseits können gleich benannte Orte falsch positive Treffer produzieren. Ein beliebtes Beispiel ist der Städtename “Alexandria”. Dieser Name taucht im Mittelmeerraum sehr häufig auf, da er auf Alexander den Großen zurückzuführen ist, der viele Städte gründete, die seinen Namen tragen. Ein anderes Beispiel sind Orte, die nach ihrer Funktion, einem Heiligen oder einer mythologischen Figur benannt sind.
Jede dieser Eigenschaften bringt eigene Probleme mit sich. In der Kombination können sich diese Probleme durch ein Fuzzy-Matching-Verfahren gegenseitig ergänzen. Eine Visualisierung von Netzen kann in diesem Zusammenhang dabei helfen, Probleme und Fehler, die bei solchen Vorhaben auftreten schnell zu identifizieren, da sich Fehler beispielsweise in großen Clustern bemerkbar machen.[101]

Zeitprobleme bei unvollendeten Datensätzen

Bei der Verbindung von Datenbanken kann es zu Problemen kommen, wenn Daten sich immer wieder ändern und definitorische Änderungen auftreten. Da statische Datenversionen nicht immer am gleichen Tag erstellt werden, können zeitliche Löcher enstehen. Das hat zur Folge, dass Teile der Daten nicht betrachtet werden können.
In der Zeit, in der ein Mapping zwischen zwei sich ändernden Datenquellen angefertigt wird, veraltet die Basisinformation der zu vernetzenden Datenquellen.
Das hängt mit der Art und der Änderungsgeschwindigkeit der Datenquellen zusammen. Im geisteswissenschaftlichen Bereich, in dem die Ressourcen eher knapp sind und sich daher wenig auf einmal ändert, ist dieses Problem wahrscheinlich überschaubar. In Datenquellen wie Wikipedia mit sehr vielen Nutzern weltweit ist diese Phänomen wahrscheinlich problematischer.

Netze und Fehler

Wie hier dargestellt, ist die Vernetzung von Datenressourcen fehleranfällig. Gerade automatische Verfahren, die vielleicht eine Präzision um die 90 Prozent erreichen, haben auf der anderen Seite das Problem, dass jeder zehnte Link, den sie erstellen, falsch ist. Selbst wenn diese Rate gesteigert werden könnte, ist die schiere Menge an falschen Links, die durch automatische Verfahren angelegt wird, für verlässliche Inferenzen aus diesen Daten fatal. Das Problem des geringen Recalls ist dabei zweitrangig, da die Vollständigkeit unschön ist und die Allgemeingültigkeit verfälscht, was aber in einem globalen, sich ständig ändernden Datennetz wahrscheinlich nie zu lösen ist. Dazu kommt, dass einige Fakten einfach nicht bekannt sind und wahrscheinlich nie bekannt sein werden.
Es sollten jedoch Verfahren genutzt werden, um visuell unterstützt systematische Fehler in Daten zu finden und zu entfernen.

4.2.4 Bereitstellung und Verfügbarkeit der Daten

Die Infrastruktur der Semantic Web- und Linked Data-Projekte ist nach der erstmaligen Erstellung der Daten die kritische Komponente. Alle Daten nutzen wenig, wenn es keine

zuverlässige Infrastruktur gibt, diese Daten nachhaltig bereitzustellen. Dazu müssen Ressourcen vorhanden sein, um dies zu gewährleisten. Um eine nachhaltige Infrastruktur für linked Open Data bereitzustellen, müssen konstant Mittel und Arbeitskraft bereitstehen. Im Folgenden werden die statische und dynamische Bereitstellung von Linked Data und auch Semantic Web Daten diskutiert.

Statische Datenquellen

Eine Möglichkeit, Linked Data zu benutzen, ist die Verwendung von festen Daten-Dumps. Im DBPedia-Projekt beispielsweise können verschiedene Daten komprimiert in diversen RDF-Formaten heruntergeladen werden. Informationen lassen sich so genau referenzieren. Der Vorteil dieses Vorgehens liegt darin, dass einfach "nur" eine Download-Plattform angeboten werden muss.
Für den Nutzer geht dabei das spontane Element verloren. Eine Verarbeitung der Daten setzt dabei den Download von einer, je nach Datenquelle, großen Menge von Daten voraus, die nicht in ihrer Gänze von Interesse ist. Es wird also zwangsweise viel Unwichtiges mit heruntergeladen. Das ist je nach Verwendungszweck, Internetanbindung und Endgerät von Nachteil.
Die Version oder wenigstens das Datum eines Dumps ermöglicht es nachzuvollziehen, welche Daten einem Verarbeitungsergebnis zugrunde liegen. Es macht Ergebnisse reproduzier- und nachvollziehbar.
Die Abfrage einer Datenbank oder eines SPARQL-Endpoints ist in dieser Hinsicht problematisch; eine Anfrage an eine dynamische Datenquelle ist im Nachhinein nicht nachvollziehbar. Das gilt auch, wenn der genaue Zeitpunkt und die Abfrageparameter bekannt sind, denn eine echte Live-Datenbank kann meistens nicht zu jedem Zeitpunkt zurückverfolgt werden.

Dynamische Abfrage

Offene, dynamische Abfragen direkt auf der Datenbasis einer relationalen Datenbank sind unüblich. Das direkte Abfragen per SQL durch Dritte wird als Schwachstelle eines Systems angesehen. In der Welt des Semantic Web ist das jedoch explizit vorgesehen. Abfragen können über öffentliche SPARQL-Endpoints getätigt werden.
Die Abfrage über SPARQL ist dabei oft einfach über das HTTP-Protokoll möglich. SPARQL ermöglicht es, beliebige Abfragen zu stellen. Dabei ist es jedoch mehr oder weniger einfach, die wirkliche Datenstruktur hinter dem Endpoint zu erkennen. Die eigentliche Struktur von RDF ist dabei nicht das Problem. Die verwendeten Ontologien können verstanden und verinnerlicht werden. Eine Ontologie zu verwenden heißt jedoch nicht in jedem Fall, den vollen Umfang einer Ontologie zu verwenden.[13]
Ein problematischer Faktor bei öffentlichen SPARQL-Endpoints sind die Bereitstellungskosten. Jede Anfrage verbraucht, je nach Komplexität, Rechenzeit. Bei vielen komplexen Anfragen kann es daher zu Problemen mit der Antwortzeit kommen. Die Auslastung liegt

[13] Siehe Abschnitt 4.1.2.

also auch in den Händen der Nutzer. Viele SPARQL-Endpoints haben dabei eingebaute Grenzen, die eingehalten werden müssen, damit Anfragen verarbeitet werden. So kann es passieren, dass nur ein begrenztes Ergebnis zurück geliefert wird. Das kann gerade bei der Suche nach Komponenten in Daten stören. Auf diese Weise können Löcher unbekannter Größe in den Daten entstehen.
Ein Vorteil der dynamischen Bereitstellung ist, dass mit wenig Aufwand explorative Analysen auf Daten gemacht werden können. Es muss vorher keine Infrastruktur für die statischen Daten aufgebaut werden, wie bei der Verwendung der Daten-Dumps. Für einen explorativen Ansatz wäre das zu aufwendig. Um einen Überblick über das Potential von Daten für die weitere Verarbeitung zu erhalten, sind diese dynamisch abfragbaren Quellen sehr praktisch.
Die Rechenzeit kann eine größere Rolle spielen, wenn ein Reasoner auf der Seite des SPARQL-Endpoints bereitgestellt wird. Durch einen Reasoner können logische Schlüsse in das Abfragesystem einfließen. Das verbraucht natürlich weitere Rechenzeit.
Oft werden die Ressourcen für Semantic Web-Anwendungen von öffentlichen Institutionen bereitgestellt. Die Dateninfrastuktur ist damit ein öffentliches Gut. Es treten in diesem System also potenziell die Probleme der Gemeingüter auf, wie sie in der "Tragedy of the Commons"[121] beschrieben wird. Die Ausbeutung durch Einzelne kann hier am Beispiel der Anfragen eines Interessenten beschrieben werden. Diese Anfragen können so komplex sein, dass andere Interessenten keine Antwort mehr vom Server bekommen, da das System beschäftigt ist.
Dabei können Restriktionen wie die Begrenzung der Suchergebnisgröße kein letztes Mittel sein. Dies ist vor allem dann wichtig, wenn nicht transparent kommuniziert wird, wie die Begrenzungen aussehen. So können auch Anfragen nach dem Absenden umgeschrieben werden, um Preformance Probleme zu vermeiden. All diese Punkte sind für die konkrete nachvollziehbare Nutzbarkeit solcher Quellen fatal.
Für eine wirkliche Expansion des Semantic Web in den kommerziellen Sektor sind daher eigene Geschäftsmodelle nötig, da hier keine Erträge durch Werbung oder Ähnliches, wie im traditionellen Web, gemacht werden können.[275]

Herausforderungen und Möglichkeiten

Wie gezeigt sind im Bereich Linked Data/Semantic Web die Herausforderungen für eine konsistente Datenverarbeitung vielfältig. Selbst wenn Daten potenziell vorhanden sein können, müssen sie an der Stelle die verwendet werden, soll vorhanden sein. Hier sollte auch beachtet werden das bei Linked Data, die Semantik der Begriffssysteme lokal variieren kann. Sollte dies gegeben sein dann stellt die passende Granularität, die konkreten Ausprägungen wie die passenden Datentypen und Skalen einen weiteren schritt zur Interoperabilität dar. Für weitere Auswertungsverfahren, wie statistische Auswertung, kann es nötig sein das Daten vollständig oder wenigstens repräsentativ vorhanden sind. Durch das Inkrementelle vorgehen und die erweiterbare Struktur gibt es große Möglichkeiten der Verarbeitung. Die verteilte Verarbeitung von Daten steht dabei jedoch durch die ungeklärte Kostenfrage zur Disposition. Um sinnvolle Applikationen erstellen zu können, ist es

daher notwendig Daten zu evaluieren und dem Entwickler zugänglich zu machen, um eine richtige Verarbeitbarkeit zu gewährleisten.

4.3 Explorativ vordringen in eine verlinkte Welt

Die Analyse von ganzen Datennetzen wie der DBpedia und noch allgemeiner der Linked Data Cloud ist nur begrenzt machbar. Die meisten globalen Analysen können nur sehr abstrakte Informationen liefern. Für die Verarbeitung und Verwendung von Daten muss ein Zusammenhang bzw. Fokus untersucht werden. Im Folgenden werden Vorgehensweisen vorgestellt die dies Leisten können und es werden dadurch einige praktische Einsichten in die Daten der DBPedia erzeugt.

4.3.1 Stichproben, Eingrenzung und Filtermöglichkeiten

Die Linked Data Cloud bietet ein unerfahrbares Meer an Daten. Doch selbst ihr Kern die DBPedia ist schon unerfahrbar. Das macht der die Betrachtung von Ausschnitten oder Stichproben interessant. Diese sollten auf eine möglichst neutrale Art beschreibbar sein. Hier ist der Unterschied zwischen einer Stichprobe und einer Auswahl oder einem Ausschnitt zu ziehen.[14] Dabei sind Stichproben theoretisch begründet.
Eine Stichprobe in der Statistik versucht eine Teilmenge zu finden, welche die Struktur des Ganzen im Kleinen repräsentiert. Dabei sollen auch Effekte der Datenerhebung und Verzerrung durch die Erhebung nivelliert werden. Eine solche Stichprobe setzt die Frage nach der Grundgesamtheit voraus.
Eine solche Stichprobe kann jedoch bei den hier untersuchten Daten kaum gezogen werden.
Ein anderer wichtiger Faktor für eine Untersuchung ist die Reproduzierbarkeit. Damit werden Daten kritisier- und überprüfbar. Auch ermöglicht eine solche Stichprobe das Aufzeigen von Entwicklung über die Zeit.
Bei der Auswahl muss ein manipulatives willkürliches Eingreifen verhindert werden. Dabei sollten die Subjekte einer Untersuchung nicht handverlesen ausgesucht werden, da so die Auswahl der Frage angepasst würde und nicht die Frage anhand der Daten geprüft würde. Es besteht also die Gefahr, eigentlich passende Daten und Entitäten herauszunehmen, um eine Frage wünschenswert zu beantworten. Eine Anpassung der Untersuchungsmenge sollte dabei mindestens eine Einschränkung der Aussage mit sich ziehen.
Bei heterogenen Datenquellen steht dabei eine Codierungsproblematik im Raum. Diese kann durch die Semantic Web- und die Linked Data-Bewegung nicht gänzlich abgebaut werden, auch wenn es durch die Semantic Web Standards versucht wurde.
Für die Untersuchung von Netzen und Zusammenhang ist es bei einer Stichprobe im Besonderen wichtig, dass Komponenten so umfassend wie möglich dargestellt werden. Das

[14] Gerade im Englischen ist die Unterscheidung schwer, da sich der Begriff “sample” hier auf eine Auswahl, einen Ausschnitt sowie eine Stichprobe beziehen kann.

Kriterium für eine Stichprobe wäre die Vollständigkeit der transitiv erreichbaren Knoten und Kanten. Das würde einen Ausschnitt erzeugen, der gleichzeitig vollständig ist.

Influence Graphs

Ein Beispiel für das Ziehen eines Ausschnitts aus DBPedia ist dabei die Auswahl einer bestimmten Beziehung oder einer Menge von Beziehungen. Hierfür werden die Typen der Subjekte und Prädikate im RDF-Tripel erst einmal ignoriert. Die daraus entstehenden Netze werden also durch ihre Beziehungen bestimmt.
Als Beispiel für dieses Vorgehen können Einflussnetzwerke, wie sie von mehreren Autoren erstellt wurden, dienen.[143, 207, 111] Bei diesen Untersuchungen geht es um zusammenhängende Komponenten von Menschen, die sich untereinander beeinflusst haben. Diese Graphen basieren dabei auf einer simplen SPARQL-Abfrage, welche mit den Properties wie `http://dbpedia.org/property/influenced` und seiner Inversion `http://dbpedia.org/property/influenced_by` arbeiten.
Einfluss geht dabei von einem Menschen auf einen anderen über. Es zeigt sich ein transitives Netz von Personen.
Es wird eine Karte der Ideenwanderung aus der Menge der Pfade von Ideen sichtbar gemacht. Dabei ist jeder einzelne Schritt für sich interessant, zeigt jedoch nicht die Ideen, die ausgetauscht wurden.
Dieses Einflussnetz ist daher interessant, da es einen Zusammenhang von den antiken Philosophen zu wichtigen Persönlichkeiten der Neuzeit zeigt.
Die Frage, was Einfluss ist und dass dies nicht klar aus der Wikipedia Info-Box-Beschreibung hervorgeht, ist ein Problem. Die Aussage von Einfluss ist dabei sehr schwach. Diese Eigenschaften lassen sich stärken, wenn der Fokus nur auf Philosophen gelegt wird. Ein Netzwerk aus Philosophen ist homogener. Die Idee ist, dass die Philosophen sich nicht in einer Weise beeinflusst haben wie sich beispielsweise Floristen oder Staatsmänner beeinflussen.
Es wird angenommen, dass ein Einfluss hier auf philosophischer Ebene stattfindet. In Zeiten, als es noch Universalgelehrte gab[15], konnten sich Philosophen auch in ihrer Rolle als Floristen austauschen.
Ein anderes Problem ist, dass ein Graph aus der DBpedia und der Wikipedia einen Bias aufgrund der kulturellen Ausrichtung beinhaltet.[111]
Das Einflussnetz in Wikipedia ist dabei ein sehr schönes Beispiel, da es so groß ist. Es bildet eine große, zusammenhängende Komponente.
Doch wie werden solche zusammenhängenden Komponenten gefunden? Welche Daten können gefunden werden und sind sie in irgendeiner Weise verwendbar? Diesen Fragen wird im Weiteren exprimentell nachgegangen.

[15] Hier kann auch der Begriff eines allgemein Gelehrten oder Ähnliches verwendet werden.

4.3.2 Exploratives Vordringen in die verlinkte Welt

Im Folgenden wird das explorative Vorgehen auf den semantischen DBPedia RDF-Daten beschrieben. Diese Daten wurden anhand von SPARQL-Befehlen definiert. Dabei werden Techniken aufgezeigt, wie ein Sample aus dem großen DBPedia Datenbestand gezogen werden kann. Dabei ist im Fokus, welche Strukturen und Verwendungszusammenhänge anhand einer Netzwerkvisualisierung und Netzwerkanalyse gefunden werden können. Am Ende werden einige praktische Probleme stehen, die aus den Daten und den dahinter liegenden Paradigmen hervorgehen.[16] Zum Einbinden des SPARQL-Endpoints wurde das Semantic Web Plugin für Gephi verwendet.[69]

Rivalität und Allianzen

Als Ausgangspunkt werden die Prädikate `http://dbpedia.org/property/rivals` und `http://dbpedia.org/property/allies` gewählt in Anlehnung an die Untersuchungen zum Einflussgraph. Diese Prädikate sind eigentlich adhoc einfach zu definieren, bei den verbundenen Entitäten handelt es sich um Rivalen und Verbündete. Eine Überlegung hierzu ist, dass der Einfluss von Bündnissen transitiven Charakter haben kann. Somit wäre eine Netzkomponente den Einfluss von Rivalitäten und Bündnissen aufeinander zu untersuchen wie in Abschnitt 2.2.1 besprochen wurde.
Da die gewählten Prädikate aus dem DBpedia-Projekt und daher aus der Wikipedia und den Info-Boxen stammen, ist das Prädikat nicht stark typisiert. Es handelt sich also nicht um eine starke Definition, wie sie durch OWL möglich wäre. Eine starke Definition in Form einer Zugehörigkeit zur Ontologie ist dabei auch in der DBpedia möglich, ist aber in der Version 3.9 noch nicht vorgenommen worden. Die Properties sind durch die Info-Boxen gegeben und dadurch nicht in einer Ontologie definiert. Insofern stellt sich die Frage, was diese Prädikate in ihrer aktiven Verwendung bedeuten und wie Allianzen und Rivalitäten durch ihren Gebrauch zu definieren wären. Dies könnte beispielsweise durch Typisierung der Entitäten am Ende der Kante festgestellt werden. Der SPARQL-Befehl 4.1 gibt eine Liste von allen Typen-Kombinationen aus, die über Rivalitäten verbunden sind. Das Vorgehen entspricht dabei der Projektion der Klassen über die konkreten Entitäten.

Listing 4.1: *SPARQL-Query zum statistischen Erfassen der Relations-Typen-Häufigkeit*

```
SELECT DISTINCT ?typea ?typeb COUNT(*) where {
  ?first ?rel  ?second.
  ?first a ?typea.
  ?second a ?typeb.
  FILTER(
  (
   ?rel = <http://dbpedia.org/property/rivals>
   || ?rel = <http://dbpedia.org/property/allies>
```

[16] Die verwendeten Daten beziehen sich auf die Ergebnisse aus dem DBpedia Live SPARQL-Endpoint vom 14.06.2014. Die generellen Erkenntnisse beziehen sich dabei auf einen viel früheren Zeitpunkt. Um eine konsistente Datengrundlage zu erhalten, sind die Analysen mit sehr wenig Abweichungen im Verlauf der Arbeit wiederholt worden.

```
  )
  && ?first != ?second
  && isIRI(?first)
  && isIRI(?second)
  && str(?first)> str(?second)
 ).
}
```

Der Befehl in Listing 4.1[17] erzeugt über 26000 Kanten und über 900 Knoten. Ein Grund hierfür ist die Struktur der Typisierung. Die Typen sind aus dem YAGO[28] Namensraum, diese sind im DBpedia SPARQL-Endpoint hinterlegt. Sie stellen einen Großteil der Ergebnisse. Die Klassen im YAGO Schema sind hierarchisiert. Diese Hierarchien werden in den Daten komplett abgebildet. Das heißt, eine Entität ist nicht nur mit der speziellsten Klasse annotiert sondern auch mit allen allgemeineren Klassen. Die transitiven Zuordnungen von Unterklassen werden alle explizit abgebildet.
Des Weiteren sind Entitäten der Klasse OWL Thing zugewiesen.[18] Das konkrete Problem ist: Die Klassen aus dem YAGO-Raum sind sehr speziell und die kombinatorischen Probleme aus der Hierarchisierung lassen sich nicht auflösen. In der Darstellung als Liste sind die Daten unüberschaubar und Gruppen lassen sich nur schwer ausmachen. Die Projektion der Daten zur Komplexitätsreduktion, wie sie in den letzten Kapiteln genutzt wurde, ist in diesem Fall unbrauchbar!
Die Dichte des resultierenden Netzes ist mit 0.047 sehr hoch und gibt damit einen verlässlichen Hinweis darauf, dass das Netz nicht zur Visualisierung brauchbar ist.

Ein struktureller Ansatz einer Definition

Da das Kategoriensystem aufgrund der explizit abgelegten, durch Inferenz erschlossenen Links nur schwer zu analysieren ist, wird deshalb ein Ansatz gewählt, der auf die unkummulierten Daten und die Verwendung der Prädikate abzielt. Der einfache transitive Graph kann dabei durch den SPARQL-Query aus Listing 4.2 erstellt werden.

Listing 4.2: *Rivalitäten und Allianzen durch eine SPARQL Query in DBpedia*

```
SELECT ?first ?second ?rel where {
  ?first ?rel  ?second.
  FILTER(
(?rel =<http://dbpedia.org/property/rivals>
||?rel =<http://dbpedia.org/property/allies> )
&& ?first != ?second
```

[17] Durch den hier dargestellten SPARQL-Befehl wird sichergestellt, dass auf der Objektposition der RDF-Tripel nur eine URI und kein Wert steht. Dafür ist im Filter-Teil der Abfrage die Funktion **isIRI** verwendet worden. Ein IRI[79] (Internationalized Resource Identifier) ist dabei eine eingeschränkte Version des URI-Konzepts, die nur auf ASCII Zeichen besteht. Diese Einschränkung ist im Weiteren für alle URIs aus dem DBPedia-Namensraum gültig.

[18] Hier ist die OWL-Basisklasse OWL Thing, von welcher jegliche Typisierung abgeleitet wird, gemeint: `http://www.w3.org/2002/07/owl#Thing`

```
&& isIRI(?first)
&& isIRI(?second) ).
}
```

Auf diese Weise werden alle Entitäten erfasst, die durch eines der beiden Prädikate verbunden sind. Die Ausgabe ist in diesem Fall eine Menge typisierter Kanten. Die Knotenidentitäten sind durch die Verwendung von URIs gegeben. Der resultierende Graph ist mit einem Verhältnis von 1032 Knoten zu 2492 und einer Dichte von 0.002 sehr viel einfacher zu verstehen und zu verarbeiten als der Klassengraph. Er besteht aus 195 einzelnen Komponenten.
Im Weiteren folgt eine Aufstellung der drei größten Komponenten aus diesem Graphen. Die Beschriftung der Knoten basiert auf dem letzten individuellen Teil der URIs aus der DBPedia.[19] Der drittgrößte Teilgraph besteht nur aus Allianzen. Dabei handelt es sich um Allianzen bewaffneter Verbände aus dem Nahen Osten, wie anhand der Bezeichner zu ermitteln ist. Dabei gibt es augenscheinlich wie in Abbildung 4.1 zu sehen zwei Cluster. Die Cluster bilden sich auf der einen Seite um Hisbollah sowie die syrische Armee auf der anderen Seite um die Israeli Defence Force (IDF). Erstaunlicherweise sind diese beiden Lager durch gemeinsame Allianzen verbunden. Bei der Recherche, warum es keine Rivalitäten gibt, wird ein spezielles Vorgehen bei der Datenauszeichnung klar. Diese Verbände und Armeen sind nicht durch "rivals" in ihrer Feindschaft abgebildet. Diese Feindschaft ist durch das Prädikat `http://dbpedia.org/property/opponents` beschrieben. Die "Zgharta Liberation Army" hat dabei in seiner Info-Box die Israeli Defence Force als Opponent und Alliierter geführt.[273] Dies scheint kein Fehler zu sein, da sie mit weiteren bewaffneten Gruppen aus dem Block um die IDF verbündet ist. Weiterhin fällt auf, dass die Betweenness Centrality und der Grad auseinanderliegen. Die Betweenness Centrality wird durch die Größe der Knoten abgebildet und der Grad eines Knotens ist durch die Farbintensität dargestellt. Die Betweenness Centrality wird genutzt, da angenommen wird, dass sich Alliierte transitiv unterstützen und zentrale Unterstützung stärkere Wichtigkeit besitzt. Außerdem bewertet die Betweeness Centrality Brückenpositionen unabhängig von deren Grad höher. Der zweitgrößte Teilgraph besteht nur aus Rivalitäts-Kanten. Die Entitäten in dieser Komponente sind Schulen. Die aufgeführten Schulen haben keine Allianz-Kanten. Die Rivalität bezieht sich auf andere Schulen und findet im Rahmen des Schulsports statt. Im Sport gibt es keine klassischen Allianzen, jedenfalls keine, die eine Zusammenarbeit bei einem gemeinsamen Spiel verursachen, da Sport außerhalb der Mannschaften kompetitiv ist.
Es handelt sich bei dieser Art der Rivalität nicht um eine transitive Beziehung. Aus diesem Grund ist hier neben dem Grad als Größe, der Clusterkoeffizient als Färbung angegeben.

[19] Um einen Knoten menschenlesbar zu dereferenzieren, kann seine Beschriftung an den Basislink `https://en.wikipedia.org/wiki/` angefügt und im Browser aufgerufen werden. Dann öffnet sich der entsprechende zugrunde liegende Wikipedia-Artikel. Dabei ist zu beachten, dass sich der Inhalt des Artikels auf aktuelle Informationen bezieht und nicht auf den Zeitpunkt zu dem diese Darstellungen erstellt wurden. Dazu muss die Version vom 14.06.2014 oder der letzen Änderung vor diesem Stichtag, herausgesucht werden.

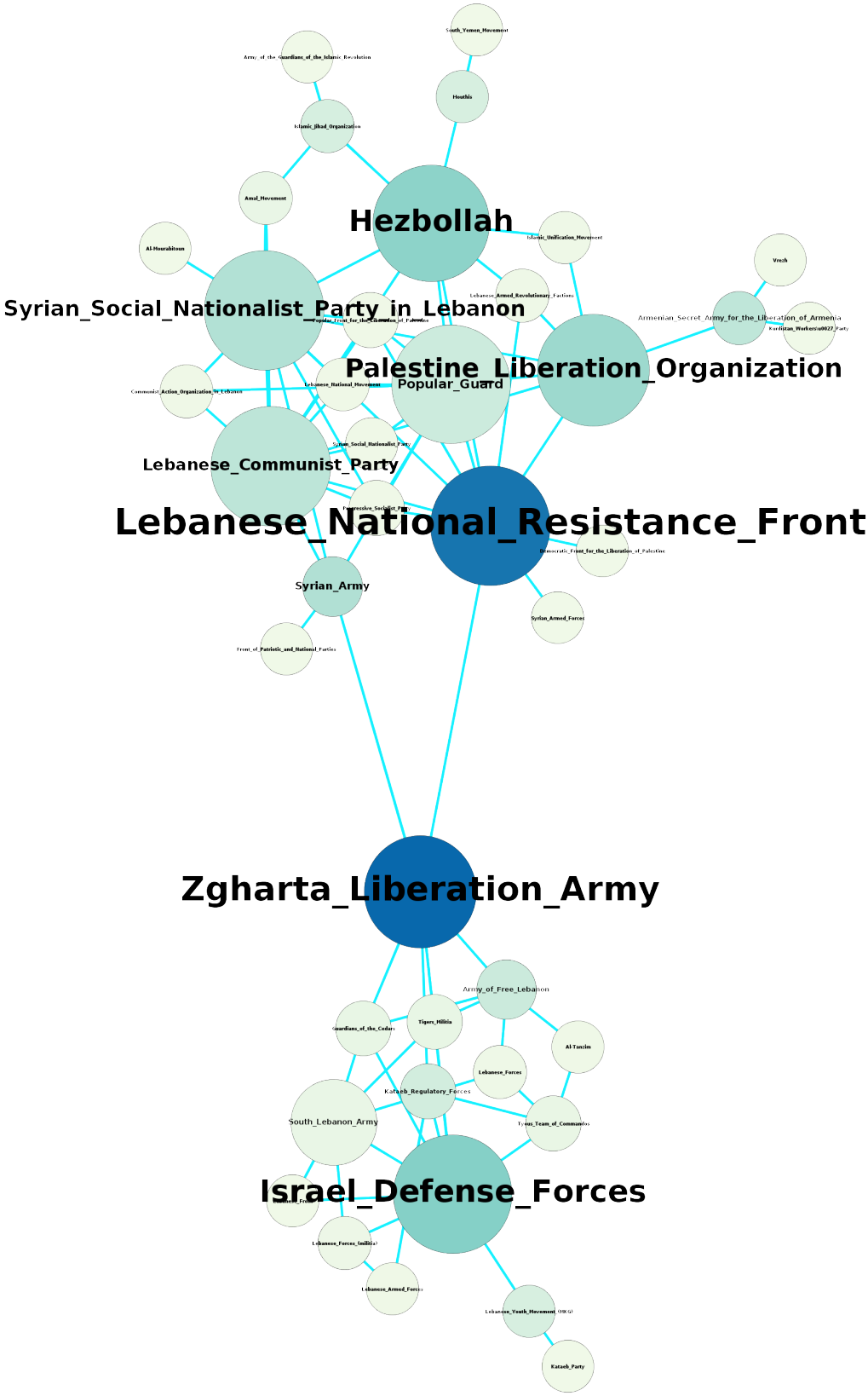


Abbildung 4.1: *Drittgrößter Teilgraph. Die Gradzahl bestimmt die Größe und die Betweenness Centrality die Farbintensität.*

Um diese Cluster bilden sich Dreiecke der Konkurrenz. Sie würden also ein Rivalitäts-Cluster bilden. Diese Cluster stehen für ein hohes, sichtbares Konkurrenzaufkommen.

Die größte Komponente besteht aus einer Ansammlung von Links, welche größtenteils kriminelle Organisationen oder andere Akteure, die mit diesen zusammenhängen, aufzeigen. Hier lassen sich nach einer Visualisierung schnell Cluster von Bündnissen erkennen. Die Komponenten sind sehr aussagekräftig. Augenscheinlich werden lokale Schauplätze sichtbar und daher auch ein Unterstützernetzwerk, das größer ist als ein Rivalitäts-Netzwerk. Die Komponenten in diesem Teilgraphen lassen sich weiter aufteilen. Der Graph aus Rivalität und Allianz kann in zwei Graphen aus Rivalitäten und Allianzen zerlegt werden. Die Knoten bleiben dabei die gleichen.

Der Allianz-Graph ist dabei der weitläufig verbundenere. Die Rivalitäten bilden eine Komponente, die 42,58 Prozent der Akteure umschließt, während die größte Komponente der Allianzen 75,21 Prozent der Akteure umschließt.

Eine Visualisierung in diesen Netzen mit verschiedenen Kanten kann nach der einen oder der anderen Kantenart erfolgen. Der hier gewählte Ansatz bezieht sich auf die Allianzen als Hauptargument für die Positionierung. Die Allianzen werden als das verbindende, positive Glieder zwischen den Akteuren gesehen. Des Weiteren gibt es mehr Allianz-Kanten, diese decken auch eine größere Komponente ab. Somit ist ein Abdriften von großen Mengen der Knoten von der größten Komponente erschwert. Eine Vorpositionierung wurde anhand der Allianz-Kanten vorgenommen, die Rivalitäts-Kanten wurden nicht mit einbezogen. Dadurch blieben einige Knoten unverbunden. Diese Knoten trieben, beim Positionieren mit einem kräfebasierten Algorithmus, von der restlichen Komponente ab. Um dies zu verhindern, wurden die Rivalitäts- und Allianzkanten mit Gewichten versehen. Die Allianzen haben höhere Gewichte bekommen und haben damit einen höheren Einfluss auf die Positionierung. Mit allen Kanten wurde dann wieder ein kräftebasierter Algorithmus auf den Graphen angewendet. Dies hatte zur Folge, dass sich die umherirrenden, freien Knoten einfangen ließen. Die Positionierung durch die Allianzen blieb dabei weitgehend erhalten. Der Einfluss der Gewichte ist auch aus den Abbildungen ablesbar. Vor allem die Details in den Abbildungen 4.5a, 4.5b verdeutlichen dies.

Abbildung 4.3 zeigt den so entstandenen Graphen. In diese Grafik sind die Rivalitäts- und die Allianzkanten in unterschiedlicher Weise eingeflossen. Die Menge der Allianzen spiegelt sich in der Größe eines Knotens wieder, umso mehr Allianzen vorhanden sind, umso größer ist ein Knoten. Die Farbe spiegelt die Menge der Rivalitäten wieder. Umso intensiver das Rot ist, umso mehr Rivalitäten hat ein Akteur.

Im rechten Teil der Grafik ist augenscheinlich, dass es mehr Kanten gibt als auf der linken Seite. Das kann daran liegen, dass dort Rivalitäten und Allianzen besser untersucht und vielleicht auch offensichtlicher sind als in anderen Teilen des Graphen. Es handelt sich dabei unten rechts um mexikanische Drogenkartelle (Abbildung 4.5a) und oben rechts um U.S. amerikanische Straßengangs (Abbildung 4.5b).

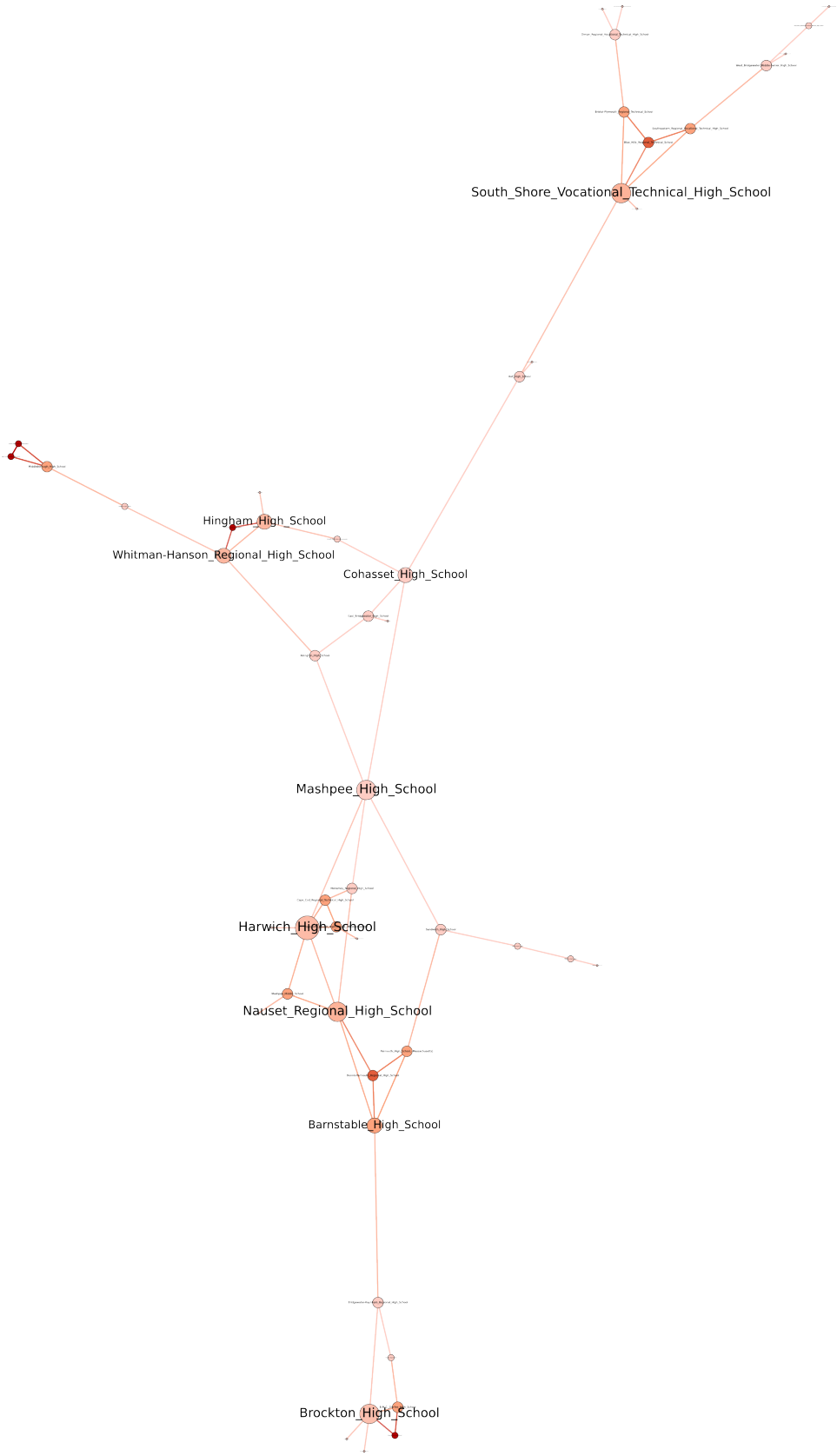


Abbildung 4.2: *Zweitgrößter Teilgraph. Die Gradzahl bestimmt die Knotengröße und der Clusterkoeffizient die Farbintensität.*

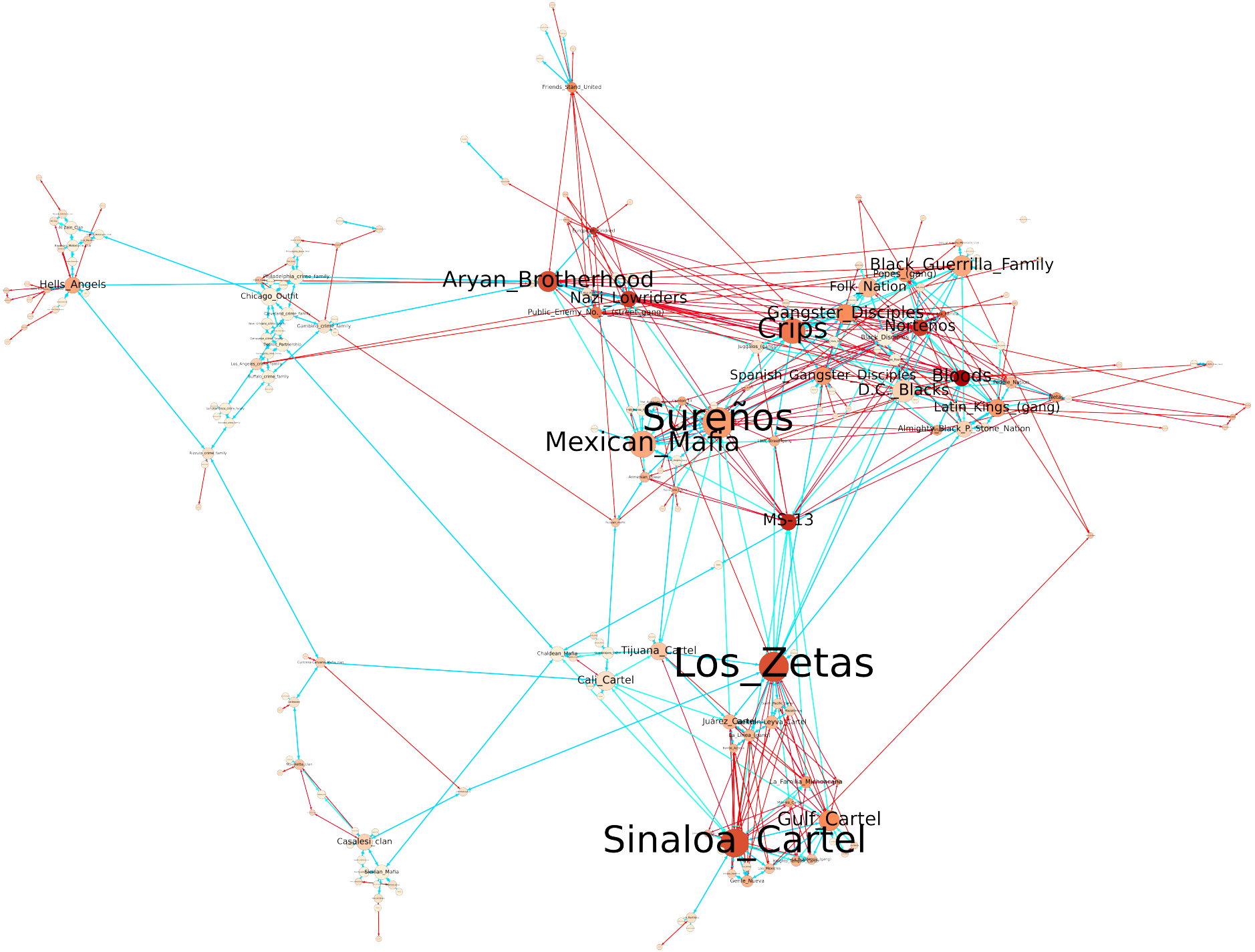


Abbildung 4.3: *Größte Komponente. Die Kantenfarben: Blau Allianzen und Rot Rivalitäten. Die Größe gibt die Unterstützung durch Allianzen an und die Farbintensität entspricht der Menge von Rivalitäten. Die Schriftgröße entspricht dem allgemeinen Grad.*

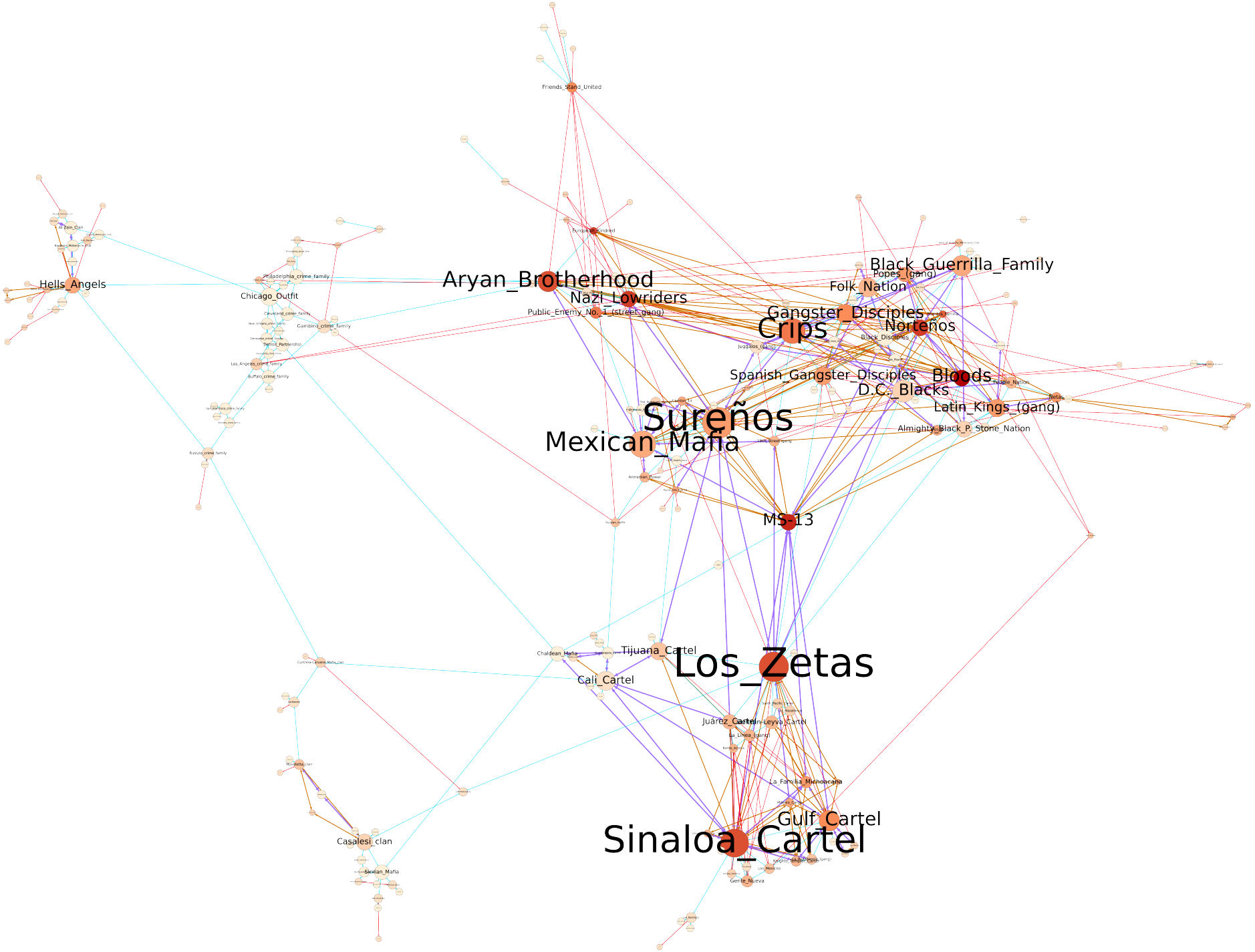
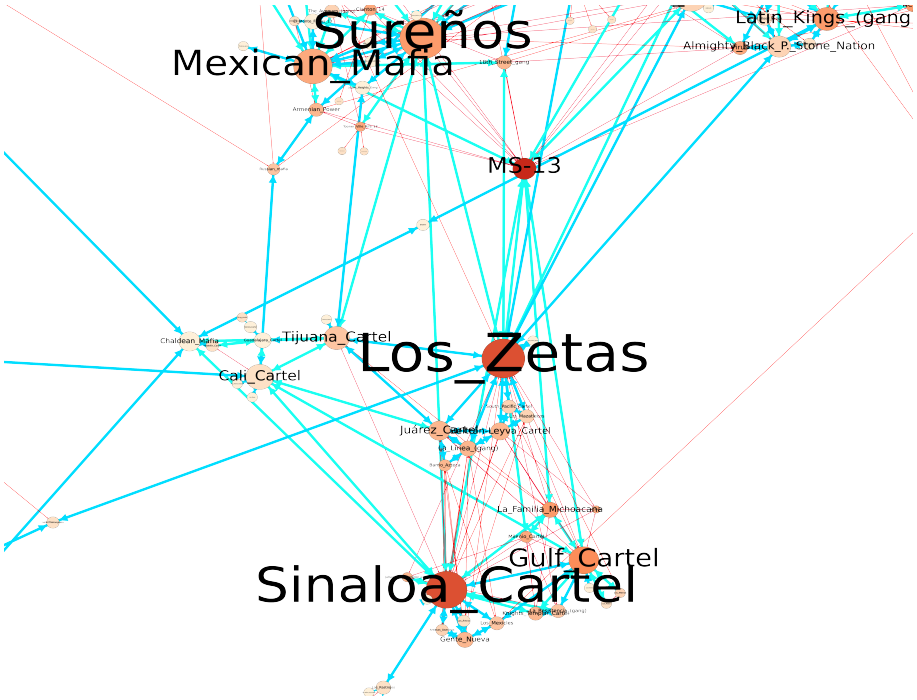
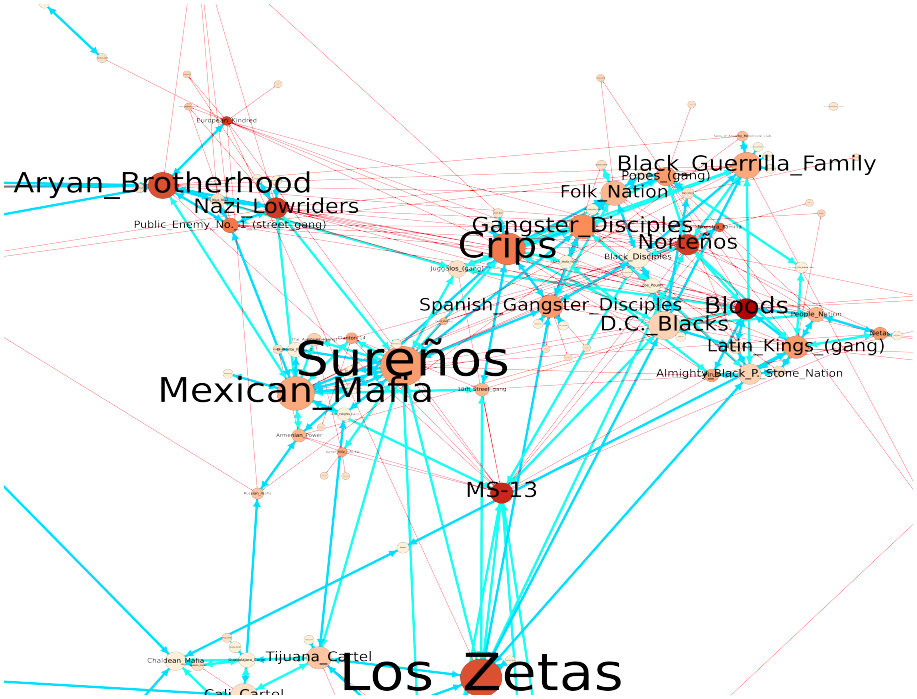


Abbildung 4.4: *Größte Komponente mit balancetheoretischen Verbindungen hervorgehoben. Orange Kanten entsprechen widersprüchlichen Rivalitäten und Lila Kanten entsprechen widersprüchlichen Allianzen.*

Abbildung 4.5: *Details aus der größten Komponente.*



(a) *Detail der größten Komponente mit Fokus auf mexikanische Kartelle (unten rechts)*



(b) *Detail der größten Komponente mit Fokus auf U.S. amerikanische Banden (oben rechts).*

Erweiterte Fragestellung

Die Grundstruktur dieser Wissenskomponente zeigt einen relevanten Ausschnitt einer Datenbasis. Doch kann durch Relationen auch eine Frage an die Datenbasis gestellt werden? Das Vorgehen bei der Frage der stabilen und instabilen Beziehungen lässt sich in SPARQL sehr gut abbilden. Hier geht es um Dreiecksbeziehungen, die einfach als stabil oder instabil bezeichnet werden können. Dies sind Teilbeziehungen, die durch die vorgegebene Struktur widersprüchlich sind. Wie in Abschnitt 2.2.1 dargestellt sind manche Beziehungskonstellationen problematisch oder instabil.
Der SPARQL-Befehl 4.3 zeigt ein Muster von instabilen Rivalitäten, die sich zwischen drei verschiedenen Gruppen bilden können. Hier würden sich zwei Feinde gegen den dritten verbünden.
Der SPARQL-Befehl 4.4 zeigt das Muster von Rivalität zwischen zwei Parteien, die aber beide mit einer dritten verbündet sind. Hier kann auch nicht gesagt werden, ob eine Allianz gekippt oder der Konflikt ausgelöst wird. Daher sind alle Kanten in einer solchen Dreierbeziehung instabil.
Diese lassen sich in diesem Fall in instabile Allianzen und instabile Rivalitäten einteilen.

Listing 4.3: *SPARQL-Query zur Konstruktion von unbalancierten Rivalitäten*

```
Construct{
?first <http://self.me/property/rivalsUnbalanced> ?second.
?second <http://self.me/property/rivalsUnbalanced> ?first.

?second <http://self.me/property/rivalsUnbalanced> ?third.
?third <http://self.me/property/rivalsUnbalanced> ?second.

?third <http://self.me/property/rivalsUnbalanced> ?first.
?first <http://self.me/property/rivalsUnbalanced> ?third.

}where{

{?first <http://dbpedia.org/property/rivals> ?second.}UNION
{?second <http://dbpedia.org/property/rivals> ?first.}

{?second <http://dbpedia.org/property/rivals> ?third.}UNION
{?third <http://dbpedia.org/property/rivals> ?second.}

{?third <http://dbpedia.org/property/rivals> ?first.}UNION
{?first <http://dbpedia.org/property/rivals> ?third.}

FILTER(
        ?first !=?second
&& ?second !=?third
&& ?second !=?third
&& isIRI(?first)
&& isIRI(?second)
&& isIRI(?third))
}
```

Listing 4.4: *SPARQL-Query zur Konstruktion von unbalancierten Allianzen*

```
Construct{
?first <http://self.me/property/rivalsUnbalanced> ?second.
?second <http://self.me/property/rivalsUnbalanced> ?first.

?second <http://self.me/property/alliesUnbalanced> ?third.
?third <http://self.me/property/alliesUnbalanced> ?second.

?third <http://self.me/property/alliesUnbalanced> ?first.
?first <http://self.me/property/alliesUnbalanced> ?third.

}where{

{?first <http://dbpedia.org/property/rivals> ?second.}UNION
{?second <http://dbpedia.org/property/rivals> ?first.}

{?second <http://dbpedia.org/property/allies> ?third.}UNION
{?third <http://dbpedia.org/property/allies> ?second.}

{?third <http://dbpedia.org/property/allies> ?first.}UNION
{?first <http://dbpedia.org/property/allies> ?third.}

FILTER(
        ?first != ?second
&& ?second != ?third
&& ?second != ?third
&& isIRI(?first)
&& isIRI(?second)
&& isIRI(?third))
}
```

Diese beiden Befehle konstruieren Meta-Informationen aus den vorhandenen Relationen. Dabei ist die Vollständigkeit der Information wichtig. Das Fehlen einer Information führt zu einem nicht gefundenen Pattern, jedoch nicht zu einem Fehler.
Die Ergebnisse dieser Analyse sind in Abbildung 4.4 visualisiert. Dabei fällt auf, dass diese widersprüchlichen Informationen vor allem in sehr stark verbundenen Komponenten auftreten. Dies kann dafür sprechen, dass die Informationen hier sehr vollständig sind oder dass diese Teile des Graphen sehr widersprüchlich sind und höheren Änderungen unterworfen sind als andere Teile. Dabei ist zu beachten, dass für eine solche Art der Untersuchung Dreiecke wichtig sind. Diese finden sich natürlich auch eher in dichten Bereichen eines Graphen.

Ewig unvollständiges Wissen

Die hier vorgenommene Auswertung hat sich nur an den wirklich identifizierten Entitäten orientiert, wie sie durch einen URI benannt werden können. Am Ende einer relationalen Kante kann dabei nicht immer nur einen URI stehen, die eine klare Entität beschreibt. Es kann auch ein schriftlicher Wert stehen. Die hier in der Auswertung beschriebenen Or-

ganisationen sind also nicht immer über direkte Links verbunden. In den Info-Boxen der Wikipedia können auch Freitext beschreibungen stehen. Dabei kann man durch das hier vorgestellte Verfahren die Entitäten zu den entsprechenden Texten feststellen. Hier wären schon in der Gesamtkomponente verbundene Entitäten eine gute Kandidatenvorauswahl für die Zuordnung von Daten. Dies wäre im Rahmen der künstlichen Intelligenz, für Algorithmen mit einer hohen Laufzeit eine mögliche Heuristik, um den Suchraum einzugrenzen.

Fazit

Die hier erstellten Darstellungen haben gezeigt, wie Semantic Web-Daten analysiert werden können und welche Probleme dabei auftreten.
Der Versuch einer Analyse der verwendeten Klassen ist nicht sehr aufschlussreich gewesen. Die hierarchische Struktur der Klassen, die explizit in den Daten verknüpft wurde, ist ein Problem. Ein kategorisiertes Subjekt wird mit allen Klassen, in denen es vorkommt, explizit verbunden. Wäre nur die spezifischste Klasse verbunden worden, wäre eine sinnvolle Analyse und Darstellung möglich gewesen. Durch das explizite Ablegen von Kategorien entsteht ein sehr dichtes Netz, das sich schlecht darstellen lässt.
Aber auch ein Vergleich auf einer Ebene der hierarchischen Klassenstruktur wäre problematisch, da hier ein Weg gefunden werden müsste, alle Klassen auf eine einheitliche Spezifitätsebene zu heben oder zu senken. Das ist mit dem Standardvokabular von SPARQL nicht beschreibbar.
Es wurde auch in weiteren Analysen gezeigt, dass die Verwendung der Prädikate nicht eindeutig ist. Eine stärkere Typisierung der Rivalitäten und Allianzen wäre aufschlussreich gewesen. Interessanterweise sind die stärkeren Typisierungen jedoch implizit in den Daten enthalten.
Bei einer Quelle wie Wikipedia haben die hier untersuchten Prädikate dazu ein Zeitproblem. Eine Rivalität muss nicht zeitlos stabil sein. Daher ist die Typisierung dieser Prädikate insofern problematisch, da sich Rivalitäten und Allianzen ändern können und auch die Abbildung ehemaliger Allianzen ihre Rechtfertigung in Wikipedia hätte. Auch die Analyse der Banden und die widersprüchlichen Allianzen unterstreichen diesen Punkt. Dabei stellt sich die Frage wie diesen komplexen Relationen in DBpedia Abgebildet werden könnten.
Ein weiteres Problem bei der Verarbeitung der Daten aus Wikipedia ist die Vermischung von Realität und Fiktion. In den Beispielen sind die “Sons of Anarchy”, eine fiktive Motorradgang aus der gleichnamigen Fernsehserie, Teil des Rivalitäts- und Allianzgraphen. Diese hat offenbar Konflikte mit der fiktionalen Version der “Hells Angels”. Das zeigt, dass eine unkritische Verwendung dieser Informationen nicht ohne Weiteres möglich ist und es Stolperfallen gibt, die in anderen Datensätzen nicht in dieser Form vorkommen.
An dieser Stelle lässt sich jedoch das Potenzial für eine Klärung der Definition der Beziehungen sehen. Eine Rivalität ist dabei nicht immer gleich. Sie wird wie hier gezeigt eher durch die Teilnehmer einer “Allianz” oder “Rivalität” und durch den transitivien Zusammenhang erzeugt. Transitiv und strukturell lassen sich die Schulrivalitäten und die Verbrecherorganisationsrivalitäten unterscheiden. Es besteht eine natürliche Entkopplung

der Netze.
Aus den hier vorliegenden Daten lassen sich einige Verwendungszwecke und Fragestellungen konstruieren.
Fragestellungen zur Vollständigkeit könnten hier nur anhand von vergleichbaren Informationen geklärt werden. Dazu müssten Daten aus offiziellen Quellen wie etwa Berichte von Strafverfolgungsbehörden extrahiert und untersucht werden. Interessant wäre es, inwiefern sich die Wissensstände unterscheiden. Dies würde zeigen, wie gut das Bild ist, das aus den öffentlichen Daten entstanden ist und wie es sich in den anderen Daten wiederfindet. Es wäre eine Art Rückkopplungsfragestellung, die auch eine Abschätzung der Verlässlichkeit von Wissen aus der Wikipedia aufzeigen könnte. Dabei stellen die gefundenen Entitäten in der Wikipedia auch einen Suchraum für Entitäten für NER-Verfahren dar. Sollten andere Textquellen erschlossen werden, könnten die Entitäten gesucht werden, die schon in dieser strukturellen Betrachtung vorkommen.
Ein weiterer Ansatzpunkt ist das Schließen aufgrund der lokalen Verwendung. Interessanterweise hat die Verwendung von Rivalität und Allianz bei den Wikipedianern keinen Anlass zum näheren Klassifizieren gegeben. Menschlich scheint der Verwendungskontext eine Spezifizierung überflüssig zu machen. Hier könnte die Datenqualität erhöht werden, wenn eine Beziehung transitiv andere Beziehungen spezifiziert. Im vorliegenden Beispiel könnten die Rivalitätskanten im Graph durch ihren Verwendungskontext entweder als sportliche Rivalität, kriminelle Rivalität oder Rivalität in einem bewaffneten Konflikt klassifiziert werden. Die Klassifikation könnte durch die Rivalitätskomponente, in der eine Verbindung steht, beschrieben werden.

4.4 Zugang, Rückkopplung und Qualität

Das Semantic Web lässt sich als komplexes Strukturmodell sehen. Nicht zuletzt durch Standards wie RDF-S und OWL wird hier eine sehr flexible Datenstruktur ermöglicht. Bei der Handhabung und Analyse kommt es jedoch zu verschiedenen Problemen auf mehreren Ebenen.

4.4.1 Zugriff und Analysierbarkeit

Für ein Netz, das potenziell automatische Agenten mit Informationen versorgen soll, sind die hier vorgestellten Standards erstaunlich schlecht zu handhaben. Der Zugriff auf SPARQL-Endpoints mit seinen Restriktionen stellt ein Hindernis für die Vollständigkeit und damit für die Netzwerkanalyse dar. Aus den Erfahrungen, die im Rahmen der Recherchen dieser Arbeit entstanden sind, machte die Linked-Data-Cloud einen sehr unvollständigen fragmentierten Eindruck, der zwar auf dem Papier schön aussieht, jedoch in der praktischen Verwendung einigen Komfort vermissen lässt. Das bezieht sich nicht nur auf das Retrieval, bei dem nur mithilfe von massivem Vorwissen und eigener Recherche Quellen übergreifende, aber leider erfolglose Datenabfragen gemacht werden können. Wenn sich zu diesen Problemen noch Unvollständigkeit und Fehler gesellen, dann ist die ganze Ideologie zum Scheitern verurteilt. Jedoch muss gesagt werden, dass die Daten schon austauschbar

sind und auch viele Werkzeuge aus diesem Umfeld lokal zuverlässig laufen und genutzt werden können.
Doch selbst wenn es einfacher wäre, Quellen übergreifende Abfragen abzusenden und es eine kritische Nutzerzahl gäbe, würden die Infrastruktur und Ausnutzung durch Einzelne schnell Lücken in die Dateninfrastruktur schlagen. Kleinere Institutionen wären wahrscheinlich mit der Bereitstellung ausreichender Ressourcen und der Pflege dieser Ressourcen überfordert, vor allem wenn gleichzeitig noch eine Primärdatenquelle wie eine Datenbank gepflegt und synchron mit semantischen Standards gehalten werden müsste.

4.4.2 Relationen und Typisierungen

Es wurde gezeigt, dass ein schwach definiertes Prädikat verschiedene Bedeutungen haben kann. Dabei zeigt sich an dem gewählten Beispiel sehr schön die Möglichkeit, dass die Abgeschlossenheit bei einer Definition helfen kann. Auch die Definition der Arten der Endpunkte einer Relation kann Aufschluss über die genauere Art der Relation enthalten. Auf diese Weise könnten verschiedene Begriffe von Rivalität und Allianz durch die Art der End- und Startpunkte einer Relation verfeinert werden.
Des Weiteren sind Beziehungen in verschiedenen Komponenten enthalten. Eine Ähnlichkeit der Beziehungen besteht durch die transitive Erreichbarkeit innerhalb eines Komplexes. Beziehungen in einer anderen Komponente sind nicht erreichbar und daher wahrscheinlich andersartig.
Auch konnte durch die Komponenten auf einen gemeinsamen, relevanten Kontext geschlossen werden.
Ebenso zeigte das Beispiel des fiktiven Motorad-Rocker-Clubs im Kontext der realen Organisation wie Fiktion bei Analysen potenziell problematisch werden kann. Daher muss gewährleistet werden, dass fiktive und echte Informationen auseinandergehalten werden, falls das für eine Betrachtung wichtig ist. Dieses Problem ist jedoch ein sehr spezifisches Problem von Quellen wie der Wikipedia, die potenziell alle Arten von Informationen enthalten können. In Hinsicht auf kulturwissenschaftliche Quellen ist dies jedoch ein wichtiger Aspekt. Beispielsweise gibt es fiktive Orte, die mit realen zusammenfallen, wie beispielsweise der Olymp als Heim der Götter und als realer Berg, an dessen Fuß sich historische Ereignisse zutrugen.

4.4.3 Zurück zur Quelle

Die hier vorgestellten Technologien und Überlegungen zeigen auch, wie stark sich Probleme des Kontexts auf der Seite von RDF-basierten Abbildungen stellen.
Quellenangaben über die Herkunft von Wissen sind generell nicht einfach zu erschließen. Die Live Version der DBPedia passt sich in Sekunden an den Wissensstand der Wikipedia an. Dabei können die Informationen von fast jedem Nutzer stammen. Dies ist auf höher akkumulierter Ebene schwer auszudrücken.
Die Wikidata-Initiative[258] versucht mit einem komplexeren Datenmodell jede einzelne Information genauer referenzierbar zu machen. Jeder Fakt ist mit einem oder mehreren

Autoren, welche auch Bots[20] sein können, annotierbar.
In der Datenhaltung kann jeder Fakt mit einem Nutzer und einer Quellenangabe versehen werden. Das ermöglicht eine Überprüfung der Fakten. Wenn Projekte, die aus Wikipedia einen Grundstock an Daten sammeln und diesen korrigieren, sollten dann diese korrigierten Daten direkt wieder in Wikipedia einfließen.
In der Wissenschaft ist es eine Vorgabe, neutral an Sachverhalte heranzugehen. Des Weiteren sollte nachvollziehbar zugearbeitet werden. Eine Manipulation der Quelldaten sollte nicht stattfinden.
Natürlich ist es nicht die Aufgabe der Forscher Wikipedia zu bearbeiten, doch wenn eine experimentelle Software Daten nutzt, die aus öffentlichen kollaborativen Quellen stammen, sollte dann ein eigener Korpus herausgenommen werden, der sich von Wikipedia absetzt oder sollten die Verbesserungen direkt auf Wikipedia eingepflegt werden?
Hier stellt sich bei Daten wie DBPedia und anderen, ähnlichen Quellen die Frage, wie eine Rückkopplung stattfinden könnte.

4.4.4 Wissen aus dem Automaten

Schon heute werden Programme benutzt, die automatisch Daten vernetzen und Fakten in Informationssysteme einpflegen. Dazu gehören einfache nachvollziehbare Regelsysteme aber auch selbstlernende Systeme.
In Wikidata sind Bots explizit vorgesehen. Diese Bots müssen Nutzern zugeordnet sein und durch die Vorlage von Beispielen durch die Community zugelassen werden. Wenn sie Schaden anrichten, werden sie gesperrt. Der Nutzer, welcher den Bot betreibt, ist für auch für seine Handlungen verantwortlich.[125] Ein automatisch erstelltes Linked Data-Web hat seine Tücken. Dabei ist gerade eine Verbindung zwischen Daten gänzlich ohne Menschen sehr problematisch. Maschinen können zwar problemlos Tausende von Verbindungen knüpfen, diese können aber auch redundant oder falsch sein, ohne dass es jemandem in der schieren Menge von Informationen auffällt.
Der Einzelne muss mittelbar oder unmittelbar an diesen Prozessen teilnehmen können. Die Technologien und Daten müssen Nutzern zugänglich und verständlich gemacht werden. Dabei ist die detailierte Ansicht wichtig, das Spezifische. Es muss aber auch eine Perspektive oberhalb der individuellen Ansicht eingenommen werden können, um größere Zusammenhänge sehen zu können.

[20] In diesem Kontext ein Programm mit einfacher Logik, das automatisch Daten, Links und Koordinaten zu Datensätzen wie Wikipedia-Artikeln hinzufügt.

Kapitel 5

Dynamische Visualisierung von Netzen

Wie in den vorherigen Kapiteln gezeigt wurde, gibt es große Menge an vernetzten Daten. Diese lassen sich meist nicht mehr erfassen und in keinem Menschenleben mehr erfahren. Dabei sind diese Daten, wie im Falle von Wikipedia, mit viel Mühe von einer großen Menge Freiwilliger erstellt worden.
Die Aufnahme in digitale Systeme und ihre Auszeichnung zeigt, wie Informationen, die früher getrennt waren, heute zu einem Großen und Ganzen zusammengetragen werden können. Wie im letzten Kapitel dargestell wird lassen sich diese Daten meist nicht sinnvoll verarbeiten, wenn ihre konkrete Struktur unbekannt ist, auch wenn die konzeptionelle Struktur festgelegt ist.
Es gibt ideelle Strukturen, wie sie durch Ontologien, Datenbank-Strukturen, XML-Schemata und kontrollierten Vokabularen geschaffen werden. Auf der anderen Seite gibt es Strukturen, die aus losen und weniger definierten Strukturen wie Bottom-up-Ontologien und Tag-Systemen bestehen. Auf der untersten Ebene stehen dabei die einfachen untypisierten Verbindungen, wie sie durch einfache Links ohne feste Semantik ausgedrückt werden.
Es muss also einen Blick auf die tatsächlichen Strukturen geschaffen werden, damit begrifflich wird, wie Daten und Formate verwendet werden.
Um den einen Sinn einer Beziehung zu verstehen, kann es notwendig sein den Kontext dieser Beziehung zu betrachten.
Das Beispiel zeigte, wie das Erkennen der genauen Semantik in Daten durch Exploration gelingen kann. Die meisten Algorithmen aus dem Feld der künstlichen Intelligenz sind auf eine solche menschliche Vorverarbeitung angewiesen. Das beinhaltet auch “saubere” Daten und eine Überwachung des Lernprozesses. Hier spielen Visualisierungen und Auswertungen eine wichtige Rolle.
Es ist jedoch mehr als fraglich, ob und wie sich eine einzige Ansicht oder Karte von diesen Datenmengen erstellen lässt. Des Weiteren ist es fraglich, ob diese Karte sich wirklich verwenden ließe. Eine Karte der Welt auf Papier gedruckt hilft nicht, wenn versucht wird, den Bäckerladen an der Ecke zu finden. Eine Weltkarte, auf der alle Straßen Fußwege,

Flugkorridore und Besitzbeziehungen dargestellt sind, lässt sich genauso wenig lesen.[1].
Das Konzept der Karte geht hier dabei über das Konzept des Abbildens von geografischen Räumen hinaus. Die Frage ist: Wie lassen sich Karten über Wissensgebiete dynamisch erstellen und was sind die Kriterien, mit denen man die für ein spezielles Problem wichtigen Pfade, Cluster und "das dazwischen" findet?
Dabei kann die Computertechnologie mehr leisten als nur eine Anordnung von Symbolen und Verbindungen bereitzustellen. Interaktive Visualisierung stellen ein gutes instrument dar, die menschlichen "Schnittstellen" anzusteuern. Computertechnologie ermöglicht es interaktive Filter anzuwenden und Hintergrundinformationen in Bruchteilen von Sekunden aus verschiedenen Datenquellen zu holen.

5.1 Dynamische Darstellungsmaschine

Im folgenden wird Anhand von Karten und dynamischen, sowie statischen Darstellugen in diesem Bereich diskutiert. Auf diese weise soll festgestellt werden wie die Vorteile einer interaktiven und statischen darstellung so Kombiniert können. Das Ziel ist eine Diskursive teilstatische Darstellung die sich an der Kartenanalogie anlehnt. Nehmen wir ein Bildnis, wie es in der Zukunft oder auch schon heute stattfinden könnte. Der Raum des Geschehens ist dabei ein Klassenzimmer. Die Lehrerin hat eine alte Schulkarte aus dem Lager geholt. Diese Karte lagerte bis vor Kurzem noch verschlossen im Raum für Arbeitsmaterialien der Schule. Im Klassenzimmer ist die Karte nun auf einen Kartenständer gehängt. Die Klasse hat Geografieunterricht. Ein Schüler wird an die Karte gerufen, um die Position der Schule auf der Karte zu zeigen. Der Schüler steht vor der Karte, er berührt die Karte genau dort, wo die Stadt mit seiner Schule liegt. Dann legt der Schüler die Finger nebeneinander und bewegt die Finger auseinander.[2] Dies wiederholt er einige Male und teilt der Lehrerin mit, dass die Karte anscheinend kaputt sei. Das Beispiel zeigt uns einige Probleme der Print-Publikation. Die Metapher der Karte funktioniert nach wie vor. Das Kind zeigt auf die Stadt, in der die Schule liegt. Ab diesem Zeitpunkt tritt ein Auswahl- und Bedienproblem auf.

5.1.1 Statisch und dynamisch

Das Kind versucht die Zoomgeste auf die Karte anzuwenden, wie es sie von seinem Mobiltelefon und der darauf befindlichen Kartenanwendung kennt. Die Zoomgeste ist der statischen Kartenleinwand jedoch fremd.
In einem Atlanten könnte in diesem Fall ein anderer, genauerer Kartenabschnitt aus dem Index gesucht werden. Alternativ könnte eine Karte der Stadt ausgewählt werden. Damit wäre eine klar referenzierbare Darstellung gewählt. Die Geste des Zoomens ist dabei auf

[1] Einen Überblick der Probleme der Kartografie zu diesem Thema bietet Skupin[234].

[2] Hier wird die zwei Finger-Zoom Geste beschrieben, wie bei Kartenanwendungen auf Geräten mit Touchscreen üblich ist. Ein Patent auf diese technische Übersetzung wurde von der Firma Apple eingereicht.[198]

den Menschen zugeschnitten. Das Auffinden von Kartenausschnitten nach einem Codierungsschema folgt einer komplexen Systematik, die auch durch einen Computer durchgeführt werden könnte.
Die Verfeinerung durch Zoomen kann im Fall der statischen Landkarte nicht stattfinden. Die Fehleinschätzung, dass die Karte kaputt sei, da sie nicht zoomen kann, zeigt die Schwächen der Karte. Die Karte macht weiterhin, was sie soll. Sie zeigt ein statisches Bild, eine Metapher des Raumes. Nur ist sie den neuen Ansprüchen an "Karten" des Kindes nicht mehr gewachsen. Gegenüber der interaktiven Kartenanwendung ist die statische Karte im Nachteil.
Der Vorteil der Schulkarte ist, dass sie nach den neuen Gesichtspunkten kaputt ist und trotzdem immer funktioniert! Das Kind übersieht, das ein gänzlich defektes Gerät in einer leeren Leinwand enden müsste. Die Karte ist also eher abgestürzt und reagiert nicht mehr.
Die digitale Karte kann, je nachdem auf welchen Daten sie beruht, sehr viel mehr sein als eine analoge Karte. Sie kann aktuell sein, also auf aktuellen Daten beruhen, während die statische Karte immer das gleiche zeigt. Die digitale Karte kann auch mehrere Maßstäbe zeigen, also in den Details, die sie darstellt, variieren. Informationen können ein- und ausgeblendet werden.[3]. Eine interaktive Karte ist mehr als die Abbildung der physischen Erdoberfläche auf einer planen Fläche mit diversen zusätzlichen Informationen.
Die digitale Karte kann aber auch das, was die statische Karte kann. Dabei ist sie die Menge aller möglichen Darstellungen. Es können also geografisch verortbare Informationen dargestellt werden und es bedarf keiner expliziten Vorauswahl, sondern die Features können on-the-fly gewählt werden.

Nachdem der Schüler weiter die Schulkarte betrachtet hat, sagt er enttäuscht: "Das ist ja nur ein Bild!" Die Lehrerin klärt das Kind auf, dass es sich um eine Karte handelt. Das Kind realisiert die Karte: "Komisch, unser Haus ist gar nicht drauf und die Straße dort kenne ich auch nicht." Die Lehrerin will den Schüler weiter auf seine Kartenkenntnisse prüfen und fragt ihn, ob er ihr den Ort zeigen kann, an dem sein Vater arbeitet. Der Schüler folgt mit den Augen einer Straße von der Karte weg und zeigt auf die Tafel neben der Karte.
Die analoge Karte verhält sich zur digitalen Karte in mehreren Aspekten wie ein Ausschnitt, der Zeit, des Ortes und der dargestellten Information. Es ist nur ein Moment aus den Möglichkeiten einer digitalen Karte. Sie könnte also die Abbildung einer Bildschirmausgabe einer digitalen Karte sein, ein Screenshot, wenn es sich um ein Rasterbild handelt oder ein Snapshot, wenn es sich um einen Zustand handelt. Dies ist im Weiteren ein wichtiger Unterschied.
Die Auswahl hinter einem statischen Screenshot kann nur sehr aufwendig rekonstruiert werden. Hier würden beispielsweise Metainformationen zum Screenshot helfen, dann ist die Rekonstruktion der Datengrundlage aber immer noch eine Frage der Rekonstruktionskunst. Um das Bild einer Karte in einen geografischen Kontext zu bringen, muss sie

[3]Beispielsweise Open Layers[190].

georeferenziert werden. Das ist mit Geographischen Informationsystemen[4] möglich, setzt aber Informationen vorraus wie: Welcher Projektion folgt die Karte? Welchen Teil der Welt bildet die Karte ab? Welche Punkte kann ich mit hoher Genauigkeit mit Koordinaten referenzieren? Diese Metainformationen sind wahrscheinlich auf einer Schulkarte vorhanden, da die Schulkarte vorbildlich ist. Die meisten dieser Informationen sind traditionell auf Karten angegeben, dazu gehören der Maßstab, die Projektion und eine Legende der Symbolbedeutung der auf der Karte verwendeten Symbole.
Ein Snapshot einer digitalen Karte ist dabei etwas Anderes und auch viel Komplexeres. Das kann alles von einer Vektordarstellung zu einer ganzen Datenbank bis hin zu einer kompletten virtuellen Maschine reichen. Letzte würde alles beinhalten das Datenmaterial und die Verarbeitungslogik. Die angezeigte Karte wäre dann ein Filter, der auf den Daten liegt.
Dieses Vorgehen kann ein Vorteil für eine kritische Betrachtung einer Argumentation sein, die aufgrund eines Kartenausschnitts getätigt wird. So kann einer Bildkritik begegnet werden. Ein Bild kann suggestiv sein, es muss aber nicht. Wenn es denn suggestiv ist, können anhand von Metadaten zu Bildern und Karten diese Suggestionen erkannt werden.
Geografische Karten sind ein recht klar standardisiertes Mittel der Visualisierung, auch wenn die draus frühe Karten schwer heutigen Anforderungen nicht mehr genügen würde.[257] Exakte geografische Methoden und globale Positionsbeschreibung sind seit der Antike bekannt.[83] Entfernungen, Winkel, Flächen lassen sich seit der Erfindung der Projektionen durch Mercator, je nach Projektionsart und Ausschnitt, geometrisch vergleichen. Heute stehen durch Satelitenbildern und Satelitenvermessung fast vollständige Daten zur Erde zu Verfügung.
Diese objektive mathematische Beschreibung von Raum ist der Grund, warum Karten heute anscheinend eher selten unter dem Verdacht der Suggestion stehen.[5] Dabei sind Karten auch immer ein Mittel der Komplexitätsreduktion. Diese Vereinfachung lässt sich nicht immer so eindeutig und endgültig beschreiben, wie es die Umsetzung in einen Algorithmus nötig machen würde. Ansicht und Projektionen werden vom Kartografen aus der Erfahrung und dem Sachverhalt nach gewählt. Die Auswahl der gezeigten Bezeichner folgt dabei keinen starren Regeln.[234]
Die Komplexitätsreduktion, die bei Kartografen per Hand vorgenommen wird, ist bei Visualisierungen und automatisch angelegten Karten die Aufgabe einer Software. Die Komplexitätsreduktion und ihre Verhältnismäßigkeit kann nur nachvollzogen werden, wenn entweder die Datengrundlage oder aber die Arbeitsweise der Software bekannt ist.
Eine der Möglichkeiten[6] eine Karte argumentativ zu verwenden ist, einen Ausschnitt zu wählen und aus der Datenlage zu argumentieren.[173] Ein Kartenausschnitt ist dabei eine bewusst getroffene Auswahl. Diese Auswahl kann schlecht gewählt oder im schlimmsten

[4] Engl.: “geographic informationsystemen”, kurz GIS.

[5] Durch online Kartensysteme lässt sich sogar visuell das Satellitenbild mit der Karte vergleichen. Damit kann selbst der letzte Zweifler, der nichts glaubt, was er nicht sieht, vom exakten Funktionieren der Kartografie überzeugt werden.

[6] Weitere Möglichkeiten finden sich bei Monmonier.[173]

Fall suggestiv sein. So kann ein Ort in einen weiteren Betrachtungskontext gesetzt werden und ein Argument kann dadurch negiert werden.
Beispielsweise die Betrachtung eines historischen Siedlungsraums. Ein Ausschnitt, der zu klein gewählt ist, könnte die Dominanz einer anderen Siedlung ausblenden. Diese Siedlung käme also erst dann zum Vorschein, wenn der Kartenausschnitt erweitert wird. Die Kritik könnte man hier auf verschiedene Arten äußern:

- Die Daten sind unvollständig. Fehlende Daten wurden nicht erhoben oder waren zu diesem Zeitpunkt nicht bekannt.
- Es wurde ein unvollständiger oder unpassender Ausschnitt gewählt.

Karten lassen anhand der Straßen und der Breite des gewachsenen Straßennetzes eine Struktur erkennen. Es ist beispielsweise ersichtlich, wie eine Stadt von Punkt A nach Punkt B durchquert werden kann. Ähnlich kann auch ein Satellitenbild verwendet werden. Dabei können verborgene Strukturen wie unterirdisch verlaufende Tunnel oder das U-Bahn-Netz nicht erfasst werden. Diese wären auf einer anderen Ebene angesiedelt. Eine einfache Analyse der sichtbaren Straßen würde also nur ein Teilbild ergeben.
Ein anderes Beispiel, das nicht so naheliegend ist, sind Flugkorridore oder Schifffahrtsrouten. Diese sind besonders, da es sich um geografische Räume handelt, die aber nur aufgrund von Aushandlungen zwischen Menschen existieren und keine physische Repräsentation haben.
Es gibt also eine Menge von Karten. Höhenkarten, Straßenkarten, satellitenbildbasierte Karten etc. Im Archiv einer Schule gibt es eine ganze Reihe von Karten, die solche Darstellungen beinhalten. Eine digitale Karte kann all das parallel sein und dabei die Informationen auf einzelnen Karten ein und ausschalten. Die Interaktivität der hier verwendeten digitalen Karte ist heutzutage faktische Gewissheit und durch diverse Programme und Onlineservices[119, 189, 30, 109] für viele Menschen auf der Welt über das Internet verfügbar. Den meisten digitalen Karten oder Serien von Karten ist gemein, dass sie eigentlich auf einer Art Grundkarte aufbauen. Diese Karte zeigt dabei geografische Einheiten wie Küstenlinien, wichtige Städte, Straßen und Ähnliches. Wichtig ist, dass alle aus dieser einen Ansicht hervorgehen Karten immer gleich bleiben. Somit wird ein Grundstein für die Darstellung verschiedenster Features gelegt. Die gleichbleibende Kartengrundlage bietet dem Betrachter Fixpunkte, an denen er sich orientieren kann.

5.1.2 Komponenten einer digitalen, interaktiven Karte

Die hier vorgestellte digitale Karte lässt sich in drei Ebenen aufteilen:

- Die Daten als Grundlage der Darstellung.
- Die Interpretationsschicht, sie macht aus den Daten eine visuelle Darstellung.
- Die Schnittstelle zum Menschen, das Interface, welches betrachtet und bedient wird.

5.1.3 Datenquellen

Der Vorteil der physischen, statischen Karte ist, dass sie aus dem Lager geholt, ohne Strom und Bildschirm, einfach funktioniert. Des Weiteren hat sie ein Datum und ist, per se, "veraltet". Der Vorteil liegt hier darin, dass jede Aussage, die aufgrund der Karte getroffen wird, auf diese eine Karte zurückführbar ist. Das Argument, das sich aus der Karte herleitet, ist dabei solange nachvollziehbar, solange die Karte existiert.
Die Lehrerin kann also ohne große Anpassung die alten Unterlagen aus dem Geografieunterricht verwenden. Bei der Verwendung einer interaktiven, immer aktuellen Karte müsste sie nachvollziehen, ob die Aussagen in ihren Unterlagen noch stimmen. Der Kontext, in dem ihre Aufgaben und Lehrmaterialien stehen, könnte sich geändert haben.
Bei interaktiven Karten ist das Nachvollziehen der Daten problematischer als bei der statischen Karte. Die Daten, aus denen die Karte erstellt wird, sind im Computer oder auf einem oder mehreren Servern hinterlegt. Die Karte ist dort keine Karte, sie besteht aus einer Beschreibung der Geo-Daten. Diese Daten werden in ein Bild übersetzt und dieses wird an den Bildschirm zur Ausgabe geschickt.
Die Darstellung hängt von der Software und im letzten Schritt von der Hardware ab.
Um eine interaktive Karte zu erzeugen, wird dynamischer Zugriff auf Daten benötigt. Das kann im einfachsten Fall durch lokalen, unveränderbaren Speicher passieren, ein Festwertspeicher [7]. Eine Karte auf dieser Grundlage wäre in dem Sinne statisch, dass die Daten nicht verändert werden können. Das bezieht sich im besonderen Maße auf das Löschen. Dieses Vorgehen würde es ermöglichen, alle erstellten Ansichten auf eine klare, nicht veränderbare Datenbasis zurückführen zu können. Dabei wäre jedoch nur eine begrenzt haltbare Karte entstanden. Bei einem Fehler in der Datenbasis könnte dieser nicht behoben werden.
Falls der Speicher doch veränderbar ist, ist eine lokale Speicherung relativ unproblematisch möglich. Sie könnte direkt in der Karte erfolgen. Damit wäre das Material, das die Karte anzeigen kann, auf die Materialien auf dem Speicher der Karte beschränkt. Die Versionen der Daten und die Verfügbarkeit wären gut vereinbar. Leiden würde hier nur die Aktualität und die endgültige Beschränktheit des Materials, da die Informationen für die Darstellungen erst auf den lokalen Speicher geschrieben werden müssen. Dazu müssten bei Korrekturen von fehlerhaften Alt-Datenbeständen diese Fehler "erhalten" bleiben, um alte Ansichten nachhaltig bereitstellen zu können.
Eine weitere Möglichkeit ist es, dass die Daten von Dritten über ein Netzwerk oder das Internet bereitgestellt werden. Beim dynamischen Nachladen von Material ist die Verwaltung der Datumsangaben komplex. Dies wird aber auf die Seite des Datenanbieters verschoben. Das nachhaltige Reproduzieren von alten Ansichten sollte dabei angeboten werden.
Die Möglichkeiten der Interaktion und Erweiterbarkeit hängt dabei davon ab, in welcher Form die Daten ausgeliefert werden, also ob Quelldaten abgefragt werden oder ob eine vorgerechnete Grafik ausgeliefert wird.

[7] Engl.: "Read-only Memory" kurz ROM.

Für diesen Ansatz ist eine Anbindung an Server mit aktuellem Kartenmaterial nötig. Auf diese hat der Nutzer nur bedingt Einfluss. Dies kann ein größeres Problem darstellen, da eine Infrastruktur mit laufenden Kosten betrieben werden muss. Dies ist kosten- und personalintensiv, da eine konstante Betreuung der Hardware sowie eine Überprüfung der Verfügbarkeit der Datenbestände durchgeführt werden muss.
Alternativ könnte Peer-to-Peer-Sharing betrieben werden, um die Daten dezentral bereitzustellen. Jeder Teilnehmer würde Daten bereitstellen. Dadurch wäre keine zentrale Datenquelle nötig und es würde auf diese Weise Ausfallsicherheit erzeugt. Es besteht jedoch die Gefahr, dass sich der Fokus auf Wissen je nach dem Alter der Daten verändert und lieber neue Daten als alte Daten geteilt werden. Neue Daten, die mehr im Gebrauch sind, haben dabei größere Aufmerksamkeit und werden deshalb öfter bereitgestellt.[200]
Ein weiteres Problem stellt in diesem Fall aber auch die Validierung der Datenquelle dar. Eine Datei ist in einem Peer-to-peer Netzwerk einfach zu verteilen, doch kann so eine Authentizität der Daten und Inhalte nicht gewährleistet sein[8]. Auch muss ein Index für die Dateien bestehen, welcher doch eine Art zentrale Infrastruktur erfordert.
Ein anderer Aspekt bezieht sich bei unfreiem Wissen auf das Digital Rights Management, im Weiteren kurz DRM. DRM ist ein sehr spezieller Fall von Verfügbarkeit, da eine künstliche Verknappung aufseiten des Bereitstellers erzeugt wird. Dieser künstliche Verknappungsmechanismus für digitale Güter muss aber auch nachhaltig instandgehalten werden. Ein Ausfallen oder ein Fehler auf Seiten des DRM Systems würde die Karte unbrauchbar machen.
DRM-Systeme, die keinen legalen Zugriff mehr registrieren können, machen die Daten und damit das ganze System unbrauchbar. Das würde die Abhängigkeit von externer Infrastruktur vergrößern, da anders als beim Server-Prinzip Datenbasen nicht einmal ohne weiteres zwischengeseichert werden könnten. Damit würde ein Ausfall der Serversysteme der Anbieter sogar bestehende abgerufene Daten unbrauchbar machen.[9]

5.1.4 Interpretationsschicht

Ein anderer Aspekt ist die Interpretierbarkeit und die Möglichkeit der Darstellung der Quelldaten.
Die Interpretation der Quelldaten und die Umsetzung dieser Interpretation sind auch einer Frage der nachhaltigen Entwicklung. Die angesprochene Reduktion der Komplexität findet an dieser Stelle statt. Dabei ist interessant, ob gleiche Quelldaten visuell anders interpretiert andere Schlüsse zulassen. Die Interpretationsunterschiede bei einer geografischen Karte sind aufgrund der standardisierten Abbildungsweise nicht in dem Maße kritisch wie sie bei nicht geographischen Darstellungen sind. Dies wird im Fall von interaktiven Gra-

[8] Die Künstlerin Madonna hat beispielsweise eine falsche Version ihres Albums “American Life” vor dem Verkaufsstart in Peer-to-Peer Netzwerke geladen. Auf den modifizierten Aufnahmen redete sie den Filesharern ins Gewissen.[160]

[9] Selbst große Anbieter von DRM-Systemen pflegen ihre Dienste nicht nachhaltig. Das kann am Fall von Zune, einem Musikdienst des Großkonzerns Microsoft, gezeigt werden. Hier wurde das Angebot eingestellt und die angebotene Hardware konnte die erworbenen Musiktitel nicht mehr abspielen und synchronisieren.[236]

phendarstellungen interessanter, da diese mehr Spielraum bei der Darstellung bieten, aber keine logisch, räumliche Grundordnung.
Als Fallstrick sind proprietäre Formate zu benennen, die nicht interpretiert werden können, da ihre Funktionsweise nicht genau nachvollziehbar ist. Das hat aber auch etwas mit der Karte an sich zu tun. Sollten Inhalte in einer Form vorliegen, die die "Karte" nicht mehr versteht, dann ist die Karte unbrauchbar. Dies kann beispielsweise bei einem Update der Karte selbst oder dem Upgrade des bereitstellenden Servers passieren.
Ein anderes gern gewähltes Mittel zur Interpretation von proprietären Daten ist eine proprietäre kompilierte Programmbibliothek. Diese ist auf eine bestimmte Computerarchitektur festgelegt, um ein Offenlegen des Quellcodes zu vermeiden und eine damit verbundene Minderung der wirtschaftliche Verwertbarkeit zu verhindern. Dies ist ein notwendiges Übel. Die Bindung an eine bestimmte Hardwarearchitektur stellt dabei auch eine nachhaltige Nutzung infrage.
Offene Software kann dagegen direkt auf den menschenlesbaren Quelltext erstellt werden. Eine Versionsverwaltung kann die Änderungen im Quelltext heraussuchen und auch eine referenzierbare Version bereitstellen. Neben der Frage nach der Nachvollziehbarkeit der Software ist die Kombination der Daten und die Wahl und Benennung des Ausschnitts aus der Datenmenge wichtig.[238] Diese Beschreibung der Datenmenge sollte, zur Vergleichbarkeit der Ergebnisse, in allen Versionen der Interpretationsschicht verstanden werden.

5.1.5 Schnittstelle

Die Schnittstelle zum Menschen ist eines der unvorhersagbarsten Elemente. Es ist beispielsweise fraglich, ob es nachhaltige Hardware geben kann.
Schneller technologischer Wandel macht nachhaltige Anschaffung von IT-Infrastruktur sehr schwierig. Es ist oft unklar, welche Systeme sich durchsetzen und wie lange Treiber für neue Systeme bereitgestellt werden. Auch ist der Austausch von Komponenten oft teurer als die Neuanschaffung.
Wie soll auf den kulturellen Wandel der Computersteuerung eingegangen werden? Ist das ganze Konzept irgendwann überholt? Auf Smartphones und Tablets haben sich Touchscreens mit Gestensteuerung durchgesetzt. Dieses Konzept existierte lange vor seinem kommerziellen Durchbruch.
Die Schulkarte ist in diesem Zusammenhang auf mehreren Ebenen eine Metapher. Auf der einen Ebene wird die geografische Karte, wie schon häufiger in dieser Arbeit, mit Graphen Darstellungen verglichen. Allgemein ist die geografische Karte ein Funktionsbildnis dafür, dass Komplexität reduziert und eine mehrdimensionale Wirklichkeit vereinfacht dargestellt wird.
Auf einer weiteren Ebene soll ein Übergang von der statischen, analogen zur digitalen, interaktiven Darstellung aufgezeigt werden. Die Vorteile der analogen Karte sollen dabei möglichst erhalten bleiben. Das Potenzial einer interaktiv-explorativen Darstellung soll jedoch, so gut es geht ausgenutzt werden. Wichtig bei einem Gerät dieser Form wäre die großflächige Darstellung. Es muss anders als die Karte eine Interaktionsfähigkeit aufwei-

sen. Diese Interaktivität muss nicht zwangsweise wie bei Touchscreens auf Tabletts oder Smartphones, Virtual-Reality Brillen oder E-Ink-Displays funktionieren.

5.1.6 Fazit

Das vorgestellte Beispiel der Schulkarte führte zu der Frage: Welche Vorteile und welche Nachteile hat diese Schulkarte im Gegensatz zu den neuen interaktiven Kartendiensten. Die Analogie zu Karten ist auch gewählt worden, um ein Gebiet einzugrenzen, in dem erfolgreich interaktiv und statisch visualisiert wird. Aber auch hier hat man nur begrenzte Möglichkeiten, die dynamische Ansicht mit einer nachhaltigen Ansicht zu verbinden.

5.2 Interaktive Graphendarstellung

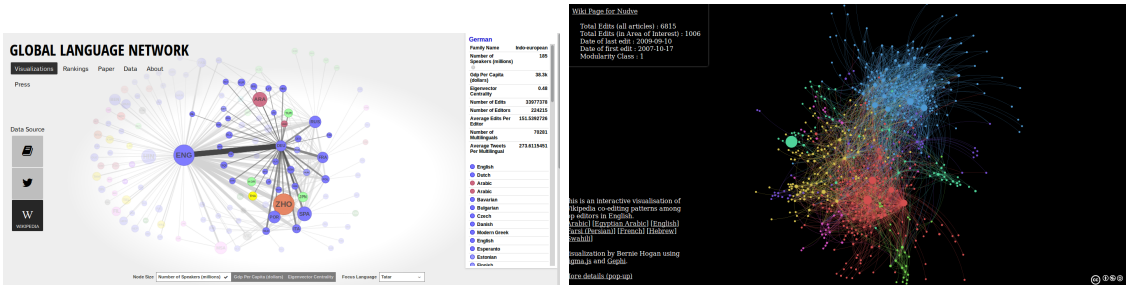
Die Darstellung von Graphen in einer interaktiven Visualisierung ist eine besondere Herausforderung. Eine solche Darstellung kann dabei nicht auf eine Analogie der Realität wie bei einer Karte zurückgreifen. Es muss ein abstrakter neuer Raum erschaffen werden. Dieser Raum kann sich dabei ähnlich verhalten wie bei einer geografischen Darstellung.
Die Netzdarstellungen, wie sie bisher besprochen wurden, beziehen sich dabei auf den zweidimensionalen Raum. Ähnlich den Karten kann man hier mit festen Symbolen arbeiten. In Abschnitt 2.3 wurde zur abstrakten Eigenschaftsmodellierung auch einige symbolsprachlich weitverbreitete Elemente vorgestellt. Diese Symbolsprache kann als eine Art Legende im Sinne von Karten dienen.

5.2.1 Interaktive Darstellungen von Wikipedia

Interaktive Darstellungen können auf statischen Daten beruhen. Das heißt, dass eine Einschränkung der Grunddaten als Vorauswahl vorliegt. Im Fall der Wikipedia beziehen sich diese Visualisierungen auf einen festen Zeitpunkt. Des Weiteren wird nur eine Auswahl an Daten betrachtet. Die genaue Herkunft der Daten ist dabei vom Retrieval und den Retrievalparametern abhängig.
Beim Projekt the Global Language Network (GLN)[106][10] wurden unter anderem die akkumulierten Verknüpfungen zwischen den verschiedenen Sprachversionen der Wikipedia untersucht. Das daraus entstehende Netz zeigt eine Übersicht über die Verlinkung der verschiedenen Sprachversionen untereinander. Dabei wird angenommen, dass Links zwischen Sprachversionen dann auftreten, wenn eine Person diese anlegt, die auch beide Sprachen spricht. Ähnliche Visualisierungen wurden auch aus Twitter-Nachrichten und Buchübersetzungen erstellt.
Die Visualisierung lässt sich hier fokussieren und es werden Daten zu den Sprachversionen und anderen Daten wie Einkommen oder Menge an Wikipedia Editoren angezeigt. [213]
Auch gibt es mehrere Projekte, die sich mit der Community-Struktur in der Wikipedia beschäftigen. Dabei handelt es sich um Co-Autorenschaft bei Wikipedia-Artikeln in verschiedenen Sprachen.

[10] Siehe Abbildung 5.1a.

Abbildung 5.1: *Interaktive Visualisierung mit statischen Daten*



(a) *Visualisierung des Global Language Networks[213]*

(b) *Visualisierung von Coautorenschaften in der Wikipedia[131]*

Die Visualisierung in Abbildung 5.1b geht auf ein Projekt von Mark Graham des Oxford Internet Institute zurück.[131] Hier wurde der Fokus des Projekts — der Mittlere und Nahe Osten sowie Nord- und Ostafrika — visualisiert. Auch wurde nur die größte Komponente visualisiert. Genaueres lässt sich aufgrund der Datengrundlage Angaben schwer reproduzieren. Die Interaktion bezieht sich hier auf Zoomen und Mouseover-Filter, der den ausgewählten Knoten und sein Umfeld hervorheben. Dazu werden Eckdaten zu den visualisierten Nutzern dargestellt, siehe Abbildung 5.1b.

5.2.2 Explorative Darstellungen von Wikipedia

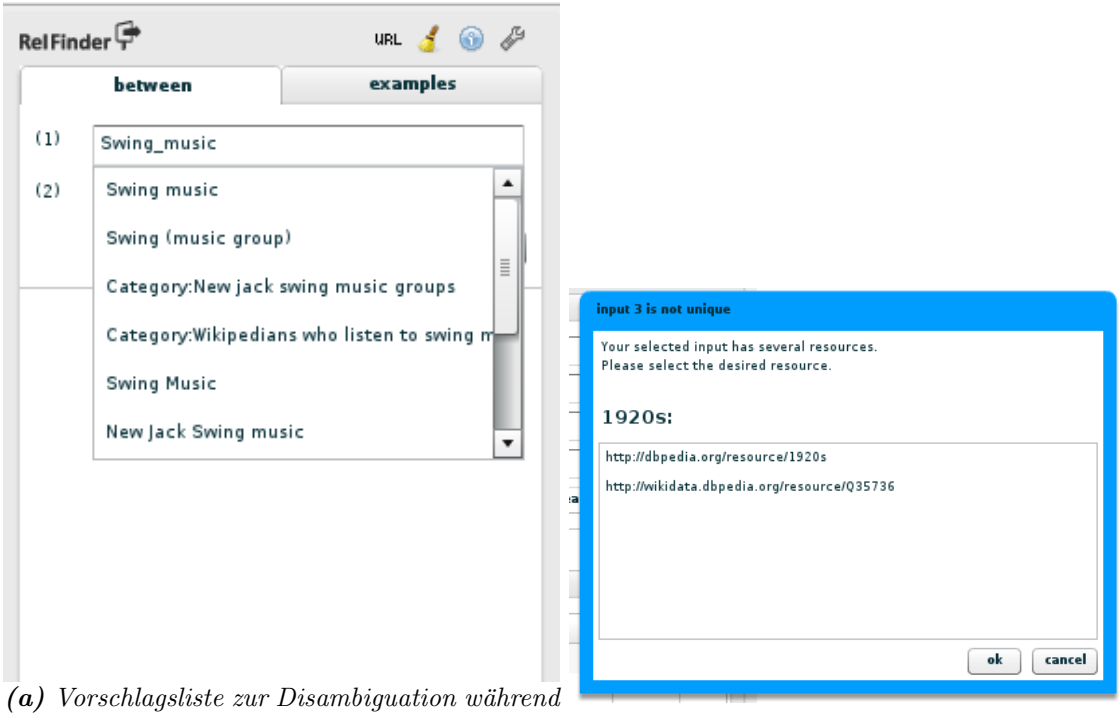
Im weiteren werden einige dynamische Darstellungen besprochen, die Daten aus Wikipedia als Netzwerke darstellen. Dabei wird ein besonderes Augenmerk auf die Reproduzierbarkeit der Ergebnisse gelegt. Auch werden die benutzten Technologien auf eine nachhaltige Nutzung hin analysiert.

Wikipedia ist ein sehr großer, frei zugänglicher und verlinkter Textkorpus. Wie beschrieben bildet das DBPedia-Projekt eine maschinenlesbare Version der Wikipedia ab. Die großen Datenmengen sind vom Menschen nicht wirklich zu erfassen. Es können immer nur kleine Teile rezipiert werden.

Um diese Daten zu verarbeiten und Anwendungen auf ihnen zu entwickeln, sind Werkzeuge wichtig, die dabei helfen herauszufinden, welche Daten überhaupt existieren. Dies steht im Gegensatz zur ideellen Form der Daten, wie sie vom Datenmodell her beschreiben wird. Bei Bottom-Up-Daten, wie sie aus Wikipedia hervorgehen, sind solche strenge formellen Strukturen nicht gegeben. Die Strukturen, die hier existieren, sind schwammig, so dass dieser Schritt eine noch größere Bedeutung für die Verarbeitung hat.

Die feste, reproduzierbare Ansicht auf Daten ist dabei ein weiteres Mittel, das in einem solchen Prozess verwendet werden kann. Die feste Ansicht schafft einen Konsens in der Darstellung. Die Reproduzierbarkeit der Darstellung aus den Daten gibt dabei ein Mittel an die Hand, einen fehlerhaft gefundenen Konsens zu erkennen und die begangenen Fehler zu benennen und zu beheben.

Abbildung 5.2: *Disambiguation von Entitäten im RelFinder*



(a) *Vorschlagsliste zur Disambiguation während der Eingabe*

(b) *Dialog zur Disambiguation*

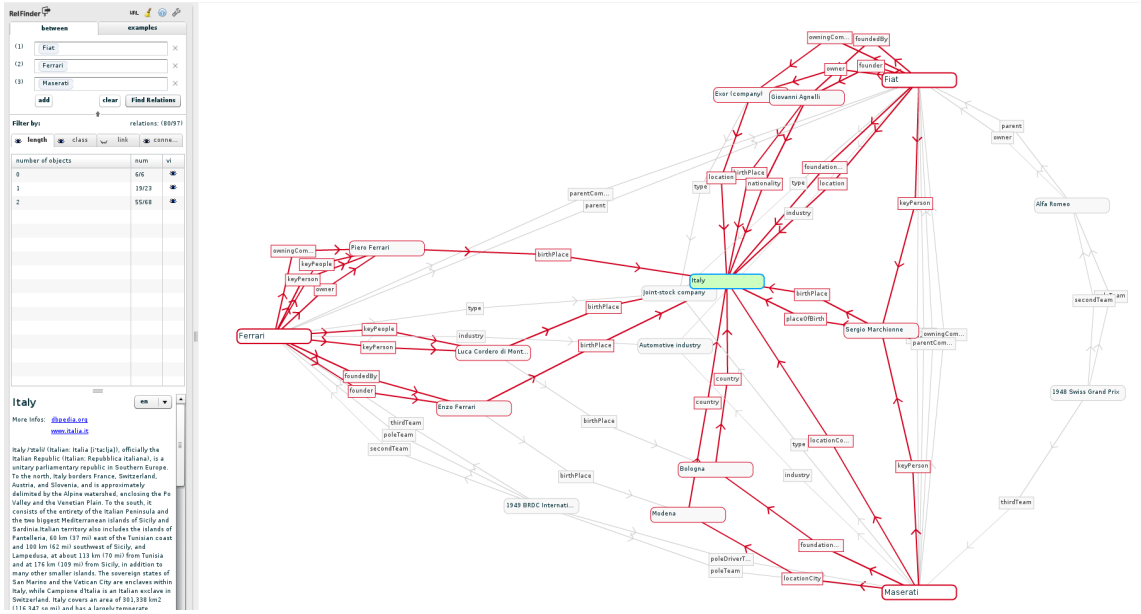
RelFinder

Der RelFinder ging direkt aus dem DBPedia Projekt hervor. Er ist ein exploratives Werkzeug um Daten aus der Linked Data-Cloud visualisieren zu lassen.[11] Der RelFinder soll als exploratives Tool für Experten dienen, um Daten aus einer speziellen Domäne zu sichten und zu bewerten.

Die Darstellung fängt mit einer Menge von Start-Entitäten an. Diese lassen sich dabei durch ein Textfeld auswählen. Da mehrere Quellen die gleichen Entitäten beschreiben oder diese wenigstens gleich heißen, ist hier eine Differenzierung notwendig. Es werden dafür bei der Eingabe verschiedene Vorschläge gemacht, um die einzelnen Entitäten auseinanderzuhalten, wie in Abbildung 5.2a. Sollte es am Ende doch zu Mehrdeutigkeit kommen, kann diese durch einen direkten Dialog aufgelöst werden wie in Abbildung 5.2b gezeigt. Die Ausgangsknoten werden von Beginn an fix positioniert, nach welchen Regeln das passiert, ist nicht genau ersichtlich. Die Darstellung im RelFinder legt konsequenterweise sehr starken Wert auf die Repräsentation der Informationen, wie sie in RDF geliefert werden. Hierzu gehören vor allem typisierte Kanten und die Richtung der Beziehungen. Diese werden durch Kombinationen von drei Knoten mit zwei Kanten abgebildet. Die Kanten fließen dabei immer in die gleiche Richtung. Der Start- und Endknoten wird durch eine URL in Subjekt und Objekt repräsentiert. Die Relation erhält einen kleineren Knoten. Die Prädikate werden anders als die Subjekte und Objekte einer Kante nicht in Einem

[11] Es gibt mehrere Versionen des RelFinders. Dabei haben auch verschiedene Versionen andere Visualisierung Techniken und Query Techniken eingebaut. Im weiteren wird sich auf die aktuelle Version unter `http://www.visualdataweb.org/relfinder.php` bezogen. Zuletzt gesehen am 13.11.2016

Abbildung 5.3: *Visualisierung des eines Beispiels des RelFinders*



dargestellt, sondern werden pro abgebildetem Triple als einzelner Knoten repräsentiert. Die Kontexte, die "zwischen" den Ausgangsknoten liegen, werden über semantische Relationen miteinander verbunden. Die dabei erstellten Pfade erreichen in den Voreinstellungen eine Tiefe von zwei Schritten. Dabei können auch längere Ketten von Verbindungen auftreten, wenn dies in den Einstellungen so gewählt wurde.
Der RelFinder kann eine Menge von SPARQL-Endpoints einbinden. Er ist nicht auf DBpedia beschränkt. Dies wird durch die flexible Struktur von RDF ermöglicht. Durch dieses Vorgehen können verschiedene Datenquellen eingebunden und parallel genutzt werden. Es werden, wie aus der Demo einfach ersichtlich, nicht nur die Relationen abgebildet. In den Einstellungen gibt es Möglichkeiten, Filter auf Attribute anzuwenden. Das ist insofern sinnvoll, da in RDF keine klare strukturelle Trennung zwischen Daten und Metadaten existiert wie in relationalen Datenbanken.
Nach der Eingabe der Startknoten, werden nacheinander die Relationen in die Visualisierung geladen. Während des Retrievals werden die Kanten positioniert. So kann die Retrievalzeit schon zum Positionieren genutzt werden. Das ist wahrscheinlich dem kräftebasierten Positionierungsansatz geschuldet. Diese Art von Positionierung ist abhängig von der Laufzeit des Algorithmus, von der Ladereihenfolge der Daten aus dem SPARQL-Endpoint und dem Einfügen in die Darstellung. Zum Start der Visualisierung sind also nicht alle Informationen verfügbar.

Die Knoten sind anklickbar und können vom Nutzer durch Drag and Drop bewegt werden. Nach der individuellen Positionierung werden die Knoten fixiert. Auf diese Weise kann der Nutzer eine beliebige Ordnung schaffen. Bei Anwählen von Knoten werden die Pfade zwischen den Ausgangsknoten und dem gewählten Knoten hervorgehoben und zeigen so die Pfade zwischen den Ausgangsknoten an. Knoten und Kanten können sich jedoch an den Rand der Darstellungsfläche verlieren.
Es existieren diverse Filter, nach der Länge der Pfade, der Klassifikation der Knoten und

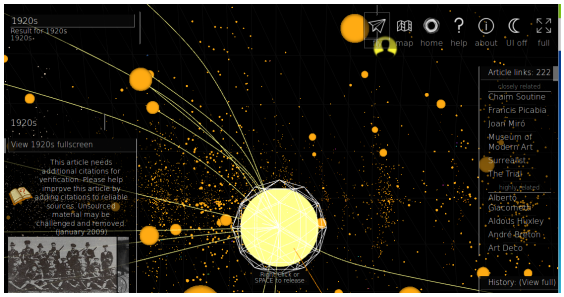
der Links. Hier zeigt sich eine Eigenschaft, die typisch für die Verarbeitung von Daten aus dem Linked Data Bereich ist. Es gibt eine Menge von Klassen und Link Typen, welche nur schwer miteinander vergleichbar sind. Zudem können diese noch aus mehreren Beschreibungsvokabularen stammen. Im Speziellen kann hier ein Dilemma beobachtet werden: Das Ausblenden einer Klasse kann Knoten ausblenden, die eigentlich auch eine andere Klasse beinhalten, die dargestellt werden soll.
Als Nutzer ist die Datenmenge für eine solche Suche kritisch. Nicht alles ist verbunden und die meisten Suchen liefern aufgrund der Unverbundenheit der Ausgangsknoten ein leeres Ergebnis.
Der Nutzer muss wissen, was er sucht. Das Problem, was sich allgemein aus Daten wie den Daten der DBpedia ergibt, ist, dass man wissen muss, was es gibt, bevor man weiß, was man finden kann. Das macht den Einstieg zum kritischen Punkt beim Arbeiten mit Linked Data. Das Ausbleiben von Ergebnissen könnte beim Nutzer zu Motivationsproblemen führen, gerade wenn mühevoll die Suchkriterien herausgesucht wurden[12]. Schnell kann das Gefühl entstehen, dass kaum Daten vorliegen.
Ein Problem der freien Positionierung ist, dass das Ausblenden von Knoten und Kanten deren Position wieder freisetzt. Nach dem Ausblenden verlieren Knoten und Kanten also ihre Position und müssen nach dem Einblenden wieder positioniert werden. Das macht eine reproduzierbare Anordnung und Weitergabe schwierig. Zwar können Suchergebnisse auch durch einen Link referenziert und somit gespeichert oder weitergegeben werden, es wird damit jedoch kein Filter und keine individuelle Positionierung weitergegeben. Damit sind Positionsaussagen wie "Links Oben“ "Mittig“ oder "Unten“ im Zusammenhang mit dieser Visualisierung nicht nutzbar. Des Weiteren kann eine Veränderung der dahinterliegenden Daten nicht genau erfassbar gemacht werden. Wenn sich die Daten im SPARQL-Endpoint ändern, dann kann ein anderes Ergebnis produziert werden.
Bei der Nutzung der Demo des RelFinders, welche auf dem SPARQL-Endpoint der Dbpedia aufsetzt, tritt das Problem auf, dass der SPARQL-Endpoint mal mehr, mal weniger ausgelastet ist. So kann es vorkommen, dass Daten nur langsam oder gar nicht geladen werden. Dies ist ein Nachteil, wie er von öffentlichen SPARQL-Endpoints ausgehen kann. Das Problem ist aber der Architektur des Semantic-Web geschuldet, wie in Abschnitt 4.2.4 beschrieben wurde.

WikiGalaxy

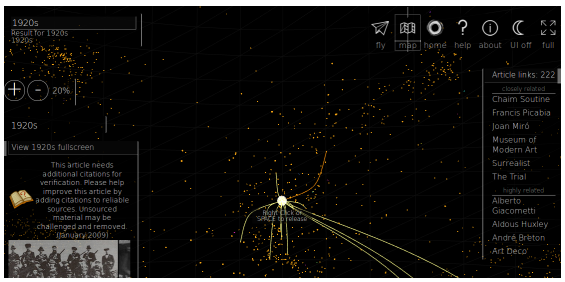
WikiGalaxy[264] bildet Wikipedia Artikel mit ihren Links ab. Die Visualisierung enthält 100.000 Wikipedia Artikel. Diese werden als Sterne in Art einer Galaxie dargestellt. In dieser Darstellung liegt der explorative Ansatz recht nah. Dabei werden grundsätzlich drei Perspektiven auf den dreidimensionalen Raum angeboten. Die detailierteste Form ist die lokale Ansicht (Abbildung 5.4a), eine rotierende Perspektive um den gewählten Artikel, die Kartenansicht (Abbildung 5.4b) zeigt eine Aufsicht auf den ausgewählten Artikel und

[12]Hier schließt der Autor von sich auf andere, doch ist es in der Argumentation einer Aufmerksamkeitsökonomie schwer, unter solchen Bedingungen Nutzer anzulocken.

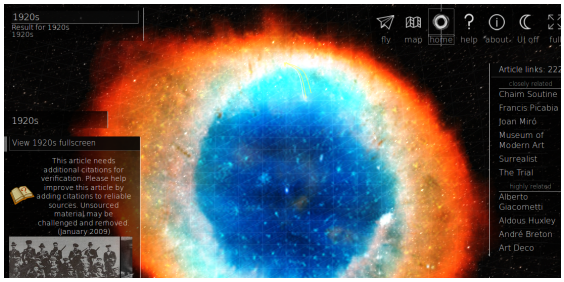
Abbildung 5.4: *Perspektiven im Wiki Galaxy Projekt*



(a) Lokale Ansicht



(b) Karten Ansicht



(c) Überblick

die Übersichtsvariante (Abbildung 5.4c), ermöglicht eine Aufsicht auf alle Daten.
Die Suche nach den Artikeln geschieht einfach über eine Eingabe, die sich augenscheinlich an den Titeln der Artikel orientiert. Diese Art der Visualisierung zeigt einen interessanten Ansatz welche Datenvisualisierungen mit webbasierten Technologien wie WebGL möglich sind. Interessanterweise wird auch hier keine wirklich dreidimensionale Ansicht gewählt, da eine Galaxie ein flacher Wirbel ist und sich in einer Art Scheibe positioniert. Das macht die Kartenaufsicht passender als es bei einer gleichmäßigen Verteilung im kompletten dreidimensionalen Raum wäre. Leider können in diesem Projekt die Ergebnisse nicht gut referenziert werden. In späteren Versionen wurde die Visualisierung dahin erweitert das die Artikel auf einer Kugel Angeordnet wurden. Das entschärft den beschriebenen Flächeneffekt.

5.3 LWMap: eine Perspektive auf Zusammenhang

Der Local Wikipedia Map[13] Service, im Folgenden kurz LWMap Service ist ein explorativer serverseitiger Dienst, der im Rahmen dieser Arbeit entstand und betrieben wurde. Er ermöglicht es Nutzern, die Verbindungen zwischen 3 bis 5 Wikipedia-Artikeln darzustellen. Dem Nutzer soll ein genereller Kontext um eine Menge von Artikeln dargestellt werden.
Der Service basiert auf den Daten der DBPedia in Version 3.9 . Die Verbindungen die angezeigt werden, stammen aus den untypisierten internationalisierten DBpedia-Page-Links. Die verbindungen representieren die Links zwischen Wikipedia Artikeln.
Zur disambiguation bei der Eingabe und der suche der Entitäten wird der Datensatz der

[13] `http://lwmap.uni-koeln.de/`

Weiterleitungen verwendet. Die typisierung der Knoten stammt aus den von DBpedia anhand der Infoboxen extrahierten Typen.
Im Weiteren wird der Service beschrieben.

5.3.1 Intention

Der LWMap Service hat die Intention, einen Ausschnitt oder ein Sample aus der Wikipedia darzustellen. Es soll einer Art Kartenauschnitt entsprechen. Dieser Auschnitt soll dem Nutzer ermöglichen, sein eigenes Wissen ins Verhältnis zum Wissen in Wikipedia zu setzen. Dabei soll das Finden von Unbekanntem und lokal Wichtigem im Vordergrund stehen.
Viele der vorgestellten explorativen Visualisierungtechniken haben den Nachteil, dass sie kaum eine Gewichtung vornehmen. Das ist einem allgemeinen Ansatz geschuldet. Das Problem eines zu generischen Ansatzes ist, dass man nicht gewichten kann. Daher ist es ein Ziel, dem Nutzer zu ermöglichen, anhand der Knotengröße neues Wissen zu präsentieren, welches nicht generell für die gesammte Wikipedia, aber im Kontext der Betrachtung wichtig ist.
Im Weiteren sollen Ergebnisse des explorativen Browsings einfach mit anderen teilbar und diskutierbar sein, ohne dass sie ihren dynamischen Charakter verlieren.

Das Finden von Interessantem

Viele statische Visualisierungen verwenden Zentralitätsmaße zum Gewichten ihrer Daten. Diese sind beim bestimmen von dominanten Akteuren wichtig. Bei speziellem Wissen und explorativen Ansätzen über Wissen ist dies jedoch nur begrenzt sinnvoll. Die Zentralitätsmaße korrelieren oft stark mit dem Grad eines Knoten. Insofern ist eine Zentralität wahrscheinlich auch mit dem vorhandenen Wissen verbunden. In einer allgemeinen Informationsstruktur ist das jedoch meistens das Offensichtliche.
Der RelFinder setzt einen starken Fokus auf typisierte Links und auf das Auffinden der Relationstypen. Die Darstellung von Semantik steht im Vordergrund. Allein aufgrund der Datengrundlage ist dies bei LWMap Service nicht der Fall.

Wiederverwertbarkeit und Referenzierbarkeit

Wiederverwendbarkeit ist ein wichtiger Punkt. Dies umfasst die gleiche Ansicht unabhängig vom Rezipienten. Daher muss jeder, der die gleiche Ansicht aufruft, auch das Gleiche sehen.
Diese Anforderung bezieht sich auf die Positionierung der Information. Die Informationen müssen in der gleichen Position zur Verfügung stehen. Die Angabe oben links sollte bei allen Nutzern die gleichen Knoten und Kanten beschreiben. Diese Erfahrung sollte auf allen Geräten Konsistent sein. Das ist bei kräftebasierten Positionierungen im Frontend meistens nicht gegeben.
Filter auf einer Visualisierung müssen beim Teilen erhalten bleiben. Wenn also Bereiche ein- und ausgeblendet werden, muss der Nutzer dies auch anderen Nutzern so weitergeben können.

Die Daten müssen nachvollziehbar sein. Daraus ergeben sich zwei weitere Bedingungen:
Die Datenbasis muss reproduzierbar sein. Es muss eine klare Version hinterlegt werden. Es muss also die Gesamtheit an Grunddaten statisch vorhanden sein.
Die Daten müssen alternativ visualisiert werden können, um einen möglicherweise irrigen Eindruck kritisieren zu können. Da die Visualisierung durch einen Algorithmus vorgenommen wird und kein Mensch eingreift, kann nicht immer die beste Ansicht erstellt werden. Um dennoch eine alternative Ansicht anfertigen zu können, müssen die Quelldaten verfügbar sein.

Technische Simplizität durch Einsatz des Web-Browsers

Komfortabel nutzbare Programme zu schreiben erfordert Zeit. Die Kernfunktionen sind dabei meist recht einfach zu implementieren. Die Implementierung von komfortabelen Oberflächen nimmt dabei einen Großteil der Implementationszeit in anspruch.
Um diese Komplexität zu verringern, wird auf die Verwendung des Web-Browsers gesetzt. Der Nutzer kann so die Features nutzen, die eh schon durch den Web-Browser bereitgestellt und die ihm mit höherer Wahrscheinlichkeit bekannt sind. Dieses Paradigma fügt sich auch gut in die Welt der Linked Data ein, da hier das HTTP-Protokoll im Zentrum steht. URIs und URLs können mit dem Browser dereferenziert werden und menschen- sowie maschinenlesbar bleiben.
Der Einsatz von Browserfunktionen ist jedoch nicht ganz unproblematisch. Browserfunktionen sind zwar weitestgehend standardisiert, lassen sich jedoch aus Missbrauchsgründen nicht direkt von einer Website aus abrufen, wie beispielsweise der Zugriff auf die Liste der besuchten Webseiten.
Auch können keine Daten zwischen verschiedenen Websites durch den Browser übertragen werden. Das wird durch die Same-Origin-Policy verhindert. Diese unterbindet das Laden von Daten, die von einer anderen Domäne kommen, um Missbrauch und das Einschleusen von Schadcode zu erschweren.

5.3.2 Abfrage

Die Abfrage des LWMap Service basiert auf Wikipedia-Artikeln bzw. DBPedia URIs. Der Nutzer muss 3-5 Artikel einstellen. Dann wird ihm aus diesen 3-5 Artikeln ein vollständiger Subgraph aus den Wikipedia Page Links gezogen.
Die Auswahl von Wikipedia-Artikeln hat dabei den Vorteil, dass die Ausgangsartikel selbstbeschreibend sind. Es müssen keine abstrakten Identifier oder URIs gesucht werden. Artikel in der Wikipedia haben selbstredende Links und URIs. Eine Entität in der DBPedia hat die gleiche URI Endung wie der entsprechende Wikipedia Artikel.
So lassen sich DBPedia-URIs direkt und eindeutig aus Wikipedia Links herleiten. Mit den URLs lassen sich einfache Abfragen erstellen und dies auch für Laien. Die meisten Laien im Internet verstehen das Konzept des Links. Wenn das Konzept der URI eingeführt wird, stellt das für den Laien eine Hürde da. Der Begriff einer URI müsste von dem der URL und dem “Link” abgegrenzt und erklärt werden. Der Unterschied ist an diesem Punkt aber

nicht ausschlaggebend. Durch die Verwendung der bekannten Begriffe “URL” und “Link” wird die Anwendung für Laien vereinfacht.

Entitätsbasierte Suche

Die Suche nach Entitäten, eindeutig definierten und abgegrenzbaren Dingen, ist anders als die Suche nach Schlüsselwörtern[14], wie sie in gängigen Suchmaschinen verwendet wird. Eine Entität ist ein mehr oder weniger konkretes, abgrenzbares Ding. Es geht über die Menge der es beschreibenden Wörter hinaus. Das wird klar, wenn es um Objekte geht. Ein Objekt kann mit vielen Worten beschrieben werden, doch lässt sich daraus keine klare Identität zu einem Objekt herstellen. Bei populären Objekten ist das ohne Weiteres möglich.
Beispielsweise bei der Nike von Samothrake. In der Wikipedia wird diese als `http://en.wikipedia.org/wiki/Winged_Victory_of_Samothrace` dargestellt. Die Arachne Datenbank bezeichnet sie als “Statue der Nike” und sie hat die URI `http://arachne.uni-koeln.de/item/objekt/14829`. Dabei muss jedoch eine Unterscheidung zu speziellen Kopien der Nike in Bonn und Berlin gemacht werden, wie sie auch in der Arachne Datenbank zu finden sind [182, 183]. Die Unterscheidung der Identitäten in Wörtern lässt sich nur sehr umständlich und missverständlich ausdrücken. Andere Identifikationsmöglichkeiten wie Bilder lassen sich von Menschen recht einfach abgleichen. Die Verarbeitung von Bildern auf dem Computer befindet sich auf einem anderen Niveau als beim Menschen. Ein die herrausorderung besteht nach dabei Bilder und deren Inhalte zu Klassifizieren. Dabei wird es einem in diesem Fall nicht weiterhelfen, da die Kopie einer Statue intentional so aussieht wie das Original.
Bei klassischen Schlüsselwortsuchen ist es dafür ohne Weiteres möglich, nach etwas Unbekanntem zu suchen. Um ein Wort zu schreiben und zu kennen, muss man nicht zwangsweise ein Bild des Konzeptes dahinter haben und wissen, was es genau bedeutet. Auch kann eine Menge von Worten in vielen verschiedenen Bedeutungen auftreten.
Die Verwendung einer Entität, wie sie durch einen Wikipedia-Artikel definiert wird, hängt auch mit der Kenntnis des Gegenstands zusammen. Es kann also Vorwissen beim Suchenden vorausgesetzt werden. Bei der Nennung einer Gruppe von Entitäten ist davon auszugehen, dass durch die intentionale Zusammenstellung ein Vorwissen über den Gesamtzusammenhang vorliegt oder wenigstens erahnt werden kann.
Die Schlüsselwortsuche steht also oft bei der Suche nach einer Entität im Vordergrund. So könnte die Suche nach “winged victory statue” bei einer Suchmaschine direkt auf den Artikel der Nike von Samothrake bei Wikipedia führen. Hier dienen die Wörter als Umschreibung der Statue, während die Identität, die durch die Bezeichnung geboten wird, selbst nur eine Beschreibung ist.
Dabei ist zu beachten, dass die Schlüsselwörter in verschiedenen Sprachen anders funktionieren. Eine Suche nach “Winged Victory Statue” liefert dabei ein anderes Ergebnis als “Geflügelte Siegesstatue”. Im Gegensatz dazu sind URIs unabhängig in der Sprache und

[14] Engl.: keywords

eineindeutig in der Beschreibung, die auch durch das Aufrufen direkt überprüft werden kann.
Bei Suchen nach Schlüsselwörtern gibt es Techniken auch ähnliche Worte zu suchen. Hier wären einige eingesetzte Techniken zu nennen:
Ein Suchbegriff kann nach seiner phonetischen Ähnlichkeit untersucht werden, es würden also alle ähnlich klingenden Worte gesucht. Die Flexion des Suchworts kann entfernt werden, wie beispielsweise bei "geflügelt" und "geflügelte" oder das Wort wird auf "Flügel" zurückgeführt. Ähnliche Schreibweisen, wie durch die Levenstein-Distanz beschrieben sind, die einen Schwellwert unterschreiten, können als valides Suchergebnis gesehen werden.
All diese Techniken sind weit verbreitet, oft implementiert und gut dokumentiert. Durch sie wird die Ergebnismenge einer Suche vergrößtert und damit ihr Recall erhöht. Diese Techniken erhöhen potenziell richtige aber auch falsche Treffer in einem Suchergebnis.
Bei Entitäten gibt es keine vergleichbaren Techniken, da die Identität einer Entität nicht auf diese Weise verhandelbar ist wie die eines Wortes.[15]
Die Suche mithilfe von Entitäten unterscheidet sich daher gravierend von der Suche mit Schlüsselwörtern und auch die Konzepte sind technisch und inhaltlich nicht austauschbar.

Entitäten benennen

Die Arbeit mit informationstechnischen Entitäten ist für Nutzer ungewohnt. Wie vorgestellt ist das Arbeiten mit Schlüsselwörtern und Suchwörtern verbreiteter. Auch wenn das Konzept einer URI in Anlehnung an eine URL oder Link den meisten Menschen verständlich ist, macht die direkte Verwendung des Konzepts in der Formulierung einer Suchanfrage einige Probleme. Ein Schlüsselwort kann direkt durch Eingabe erfolgen. Eine Entität muss gesucht und benannt werden. Dann muss sie ausgewählt und mit anderen kombiniert werden. In Form einer URI würde das einen deutlich größeren Eingabeaufwand bedeuten.
Um eine komfortable entitätenbasierte Eingabe zu ermöglichen, wurden verschiedene eingabe Szenarien berücksichtigt.

Direkte Eingabe
: Es wird ein Suchstring, der ein bestimmtes Ding oder einen bestimmten Fakt beschreibt, direkt eingegeben. Dies wird dann zu einem URI umgeformt.

Copy und Paste
: Es wird ein Link direkt aus der Wikipedia oder der Adresszeile des Browsers kopiert.

Überschrift, Benennung
: Es wird ein Bezeichner eines Links oder eine Überschrift eines Artikels aus der Wikipedia kopiert.

[15] Hier können Netze über das Same-As Prädikat aus RDFs hergestellt werden und problematische Language Links aufgestellt werden, doch sind es intentionale Kriterien, die mit Hilfe von Hintergrundwissen erzeugt werden. Es sind keine Verfahren, die sich direkt algorithmisch auf den Bezeichner anwenden lassen.

Remix

Das Suchinterface des LWMap Service stellt einige Elemente bereit, die es dem Nutzer ermöglichen, seine vorherigen Suchen oder die Beispiele neu zu rekombinieren.

Drag und Drop aus der Browser-History

Es wird ein Link Per Drag und Drop aus der History des Browsers genommen.

Drag und Drop aus dem Interface

Im Interface kann der Nutzer per Drag and Drop URIs aus den vorherigen Suchen und den bereitgestellten Beispielen auswählen.

Die verschiedenen Szenarien beruhen auf intelligenten Eingabefeldern. Die Eingabefelder sollen so eindeutig wie möglich arbeiten. Das ist mit Kosten bei der Bedienbarkeit verbunden. Bei textbasierten Suchen oft genutzte Verfahren wie Stemming und Levensteindistanz wären in einem solchen Prozess hinderlich, da ähnliche Entitäten gesucht würden und somit eine klare Zuordnung verloren gehen könnte. Um verschiedene Schreibweisen und andere Ungenauigkeiten bei der manuellen Eingabe abzufangen, werden die original DBpedia-Redirects, also die Weiterleitungen von Schreibweisen in den URLs, verwendet. Hier werden die bereits durch das DBPedia-Projekt bereitgestellten Daten wiederverwendet. Auf diese Weise können Erfahrungswerte aus der Ursprungsdatenquelle verwendet werden. Eine ungenaue Heuristik in Form von Vergleichsverfahren wird dadurch vermieden.

Identifizierbarkeit von Entitäten

Netzwerk-Analyse im klassischen Sinne dient oft als Mittel zum Finden von Gruppen oder Clustern und Akteuren in auffälligen Positionen. Die hier angestrebte Positionierung soll wichtige seltene Fakten in einem Kontext finden helfen. Sie geht über das Einzelne hinaus. Bei einer Datengrundlage wie Wikipedia Links ist es wichtig, wie man es dem Nutzer ermöglicht, sich einen Überblick zu verschaffen. Die individuelle Identität eines Knotens darf nicht verloren gehen. In der Soziologie geht es eher um das Identifizieren von Gruppen oder einer Klasse von Individuen an wichtigen Positionen im Verhältnis zu ihren Eigenschaften. Dabei steht gerade das klar benennbare Individuum nicht im Mittelpunkt. Im Gegenteil, es wird aus guten Gründen anonymisiert und ausgeblendet. Bei wichtigen Kleingruppen-Analysen kann man einen Code oder ein Symbol verwenden, da die Individualität in der Abgrenzung wichtig ist, aber nicht zur Identifikation der konkreten Person genutzt werden soll.
In anderen Bereichen, wie der historischen Netzwerkforschung, ist es anders. Hier stehen die Personen klar im Vordergrund, da es hier keine Persönlichkeitsrechte mehr zu schützen gibt.
Ein klarer Vorteil einer Visualisierung von Gruppen oder Typen ist die Eingrenzung der Identität auf wenige Merkmale. Diese können näher beschrieben werden, müssen aber nicht jedes Ding einzeln identifizierbar machen. Viele teilen sich eine "Identität". Diese Identität kann durch ein einheitliches Zeichen wie eine Farbe oder eine Form und eine Legende

mit der Bedeutung der Zeichen erfolgen. Bei der Darstellung von einzelnen Identitäten muss der Einzelne identifizierbar sein. Oft sind jedoch Beschreibungen zu lang und daher schlecht darstellbar. Auch überlagern sich die Bezeichner mit höherer Wahrscheinlichkeit, je länger sie sind.
Daher werden nur die wichtigsten Bezeichner und die der Ausgangsartikel angezeigt. Durch das Zoomen auf die Visualisierung werden diese kenntlich gemacht. Um einen unwichtigen und allgemeinen Artikel zu identifizieren, muss herangezoomt oder fokussiert werden. Dies ist eine Standardfunktion, wie sie bei interaktiven Karten zu finden ist. Hier ist jedoch anders als bei Städten die Wichtigkeit nicht einfach zu bestimmen. In der Visualisierung bezieht sich diese Darstellung auf den Zoom-Faktor und auf die Auswahl durch Filter. Die Bezeichner von ausgewählten Knoten werden daher auf jeden Fall dargestellt.
Im Rahmen der Wikipedia-Artikel kann die Identität über den individuellen Teil der URI erfasst werden. Diese entspricht meist dem Titel des Wikipedia-Artikels, weist in bestimmten Fällen jedoch auch Spezifikationen in Klammern hinter dem Namen auf. Hier wird der Vorteil eines von Menschen aufbereiteten Korpus genutzt. Die Identität sollte dadurch ausreichend einfach beschrieben sein, ohne dass eine Verwechselungsgefahr besteht.
Eine weitere Möglichkeit den Titel zu verfeinern ist den Kontext anzuzeigen. Interaktiv wird beim ersten klick auf einen Knoten der Kontext, also alle direkt verbundenen Artikel, hervorgehoben. Alle anderen Knoten werden an dieser Stelle ausgeblendet. Weiterhin werden die Identifier bzw. Labels hervorgehoben.[16]
Falls dies nicht ausreicht, um die Identität festzustellen, ist das Aufrufen des Artikels nötig. Dies geschieht beim zweiten Klick auf den ausgewählten, hervorgehobenen Artikel. Es könnten auch weitergehende Beschreibungen in der Datei selbst abgelegt werden. Rechtlich stellt dies aufgrund der Lizenz, unter der die Daten stehen, kein Hindernis dar. Wenn aber diese weitergehende Beschreibung direkt in die Quelldatei eingefügt wird, dann nimmt das, je nach Menge der Knoten, eine signifikante Menge an Speicher ein. Auch wäre das Hinterlegen der Daten auf dem Server hinter einer Abfrageschnittstelle möglich. Diese Funktion wäre jedoch beim Ausfall des Retrievals und Datenbanken nicht mehr vorhanden.

5.3.3 Retrieval

Eine Abfrage erstellt einen kompletten Graphen. Dabei sind die Ausgangsknoten, die als Suchparameter fungieren, maßgeblich für die Auswahl des Datenausschnitts verantwortlich. Es werden alle Artikel in das Retrieval einbezogen. Es gibt im entstehenden Datenmodell zwei Arten von Knoten und drei Arten von Kanten.

Knoten-Typen

Ausgangs-Knoten

Dies sind die Artikel, welche bei der Abfrage angegeben werden. Im weiteren werden sie als Knoten des Typs 1 oder Ausgangsknoten bzw. Ausgangsartikel bezeichnet.

[16] Dies ist eine Funktion, die auch im Standard von Gephi und SigmaJS enthalten ist, wird jedoch im Programm weiter modifiziert.

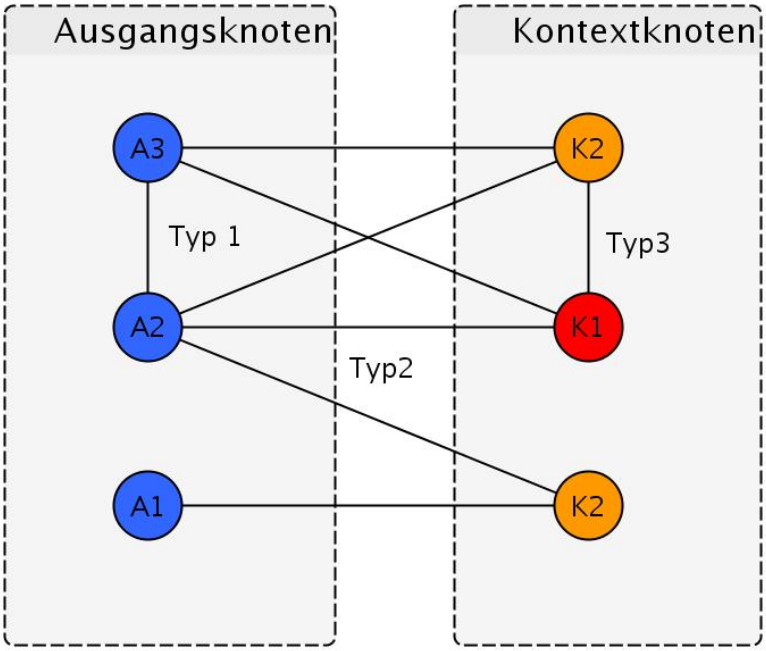


Abbildung 5.5: *Darstellung der Retrieval-Komponenten mit den entsprechenden Knoten und Kanten Beschriftungen.*

Kontext-Knoten

Dies sind die Artikel, welche direkt zwischen den Ausgangsartikeln liegen. Sie sind immer einen Schritt von den Ausgangsartikeln entfernt. Im Weiteren werden sie als Knoten des Typs 2, Kontextknoten bzw. Kontextartikel bezeichnet.

Kanten-Typen

Es gibt drei Kanten-Typen. Die Klassifikation dieser Typen lässt sich durch die Knotentypen erklären, die sie verbinden.

Kanten zwischen den Ausgangsknoten

Diese Kanten beschreiben den direkten Zusammenhang der Artikel. Alle diese Links sind auch beim direkten Browsen der Ausgangsartikel ersichtlich. Im Weiteren werden sie als Kanten Typ 1 bezeichnet.

Kanten zwischen den Ausgangsknoten und Kontextknoten

Das sind die Kanten zwischen den Knoten des Typs 1 und Typs 2. Sie bilden einen adhoc bipartiten Graphen zwischen den Ausgangsartikeln und den Kontextartikeln. Im Weiteren werden sie als Kanten Typ 2 bezeichnet.

Kanten zwischen den Kontextknoten

Diese Links beschreiben die Verbindung der Kontexte untereinander. Im Weiteren werden sie als Kanten Typ 3 bezeichnet.

Retrievalverlauf

Das Retrieval holt alle Knoten, die mit den Ausgangsknoten verbunden sind aus der Datenquelle. Die Richtung der Verbindung wird nicht mit in Betracht gezogen. Alle Verbindungen werden in diesem Schritt gleich verarbeitet, eingehende wie ausgehende Verbindungen. Die entstehenden Knoten werden gezählt. Dieses Vorgehen kann und wird parallelisiert ausgeführt, um eine bessere Performance zu gewährleisten.
Wenn ein Artikel also mit zwei Ausgangsknoten verbunden ist, dann wurde er auch zwei Mal gezählt. Alles was nur mit einem der Ausgangsknoten verbunden ist, wird verworfen. Die restlichen Knoten bilden die Kontextknoten. Gleichfalls werden bei dieser Operation die Zuordnungen zu den Ausgangsknoten gespeichert. Das liefert die Kanten des Typs 2.
Im nächsten, dem aufwendigsten Schritt werden alle Kanten des Typs 3 erstellt. Diese erhält man, wenn alle ausgehenden Links der Kontextknoten aus der Datenquelle geholt werden. Beim anschließenden Filtern bleiben nur die Kanten erhalten, die als Ziel- und Quellknoten Kontextknoten haben. Auch dieser Vorgang lässt sich parallelisieren. Der Schritt stellt ein wichtiges Alleinstellungsmerkmal gegenüber den anderen Visualisierungen im Wikikontext, wie sie in Abschnitt 5.2 vorgestellt wurden, dar.
Im Retrieval werden on-the-fly weitere Werte festgehalten. Hierzu gehört auch der In- und Out-Degree im Ausgangsgraphen. Diese Zahlen werden später zur Bewertung der Knoten verwendet.
An dieser Stelle ist es möglich, auch andere Attribute aus dem DBpedia Datensatz dazu zu laden wie beispielsweise die Klassen aus der DBPedia Ontologie. Dabei ist zu beachten, dass hier nicht das RDF-Datenmodell verwendet wird. Die Klassen werden nicht als eigenständige Knoten behandelt, sondern als Attribute mit mehreren Werten, Attribute die nicht nur eine Ausprägung haben können, sondern viele. Dieses Vorgehen bietet andere Ansätze zur Verwendung dieser Werte wie z.B. durch facettiertes Filtern.

Filter

Filter im Retrieval treten in zwei Fällen auf:
1. Es gibt zu viele Kontextkonten zwischen zwei Ausgangsknoten. Diese werden dann aus pragmatischen Gründen herausgefiltert.
Inhaltlich lässt sich dies mit der Überfrachtung einzelner Kontextpositionen begründen. Es wird nach dem Speziellen gesucht, daher ist es hinderlich, wenn zwei Ausgangsknoten so viele Kontexte aufbauen, dass diese überlagert werden. Eine große Menge von Kontexten auf einer Position bedeutet auch immer einen starken generellen Zusammenhang zwischen zwei Knoten bzw. Artikeln. Gerade bei Knoten, die viel miteinander zu tun haben und auch generell stark verlinkt sind, kann dies zu sehr großen Schnittmengen führen.
Dadurch ist die Übersichtlichkeit in der Visualisierung gefährdet. Die allgemein stark zusammenhängenden und stark verlinkten Knoten nehmen dann sehr viel Platz ein. Eine Bevorzugung der Suchbegriffe soll vermieden werden. Ein Grund ist die Retrievalzeit, da eine Vielzahl von Kontextknoten das Retrieval der Kanten vom Typ 3 verlängert.
Beim Filtern bleiben die Kernknoten erhalten. Es werden zuerst die Knoten gefiltert, die stark in der Peripherie liegen, also Kontexte die mit zwei Ausgangsknoten zusam-

menhängen. Dabei muss zusätzlich ein Grenzwert in der Gesamtzahl der Knoten im Graphen überschritten werden. Das bedeutet, dass diese Art des Filterns erst eintritt, wenn insgesamt zu viele Knoten aus dem Retrieval kommen.
2. Artikel wie Listen und Indizes werden herausgefiltert. Diese Listen sind meist nicht vollständig. Dabei haben sie einen aufzählenden Charakter und verstärken den Bias. Auch beschreiben sie oft nur nach Alphabet geordnet Klassen von Dingen und das nicht zwangsweise unvollständig.[48] Sie haben meist lange Titel, da sie mit "list of" oder Ähnlichem beginnen. Listen werden daher gefiltert.

5.3.4 Gewichtung

Gewichtungen in einem Darstellungssystem haben meistens einen abstrahierenden Charakter. Das individuelle Merkmal wird dabei durch einen abstrakten Wert verkörpert. Gewichtungen können also nur immer mit einer Zielintention definiert werden. Das gilt im Besonderen für kategoriale Daten.
Die Frage und die Intention bestimmt dabei die Gewichtung einzelner Elemente, hier Knoten und Kanten. Da in der Visualisierungsphase kräftebasierte Algorithmen verwendet werden, bestimmt die Gewichtung der Knoten und Kanten die daraus hervorgehende visuelle Repräsentation. Die Gewichte der Kanten zeigen an, wie stark der Einfluss der Kanten auf das physikalische Modell ist. Die Größe der Knoten repräsentiert ihre Wichtigkeit und damit die Gewichtung und den Platz, den sie sich dadurch einnehmen.

Lokale Bewertung

Im vorangegangenen Kapitel wurde eine theoretische Betrachtung und der Abgrenzung zwischen Datenbank- und Semantik-Web-Welt gewählt, die offene oder geschlossene Welt Annahme[17]. Dabei wird mit der Vollständigkeit des Wissens argumentiert. Dies hat praktische Konsequenzen für das Treffen von Negativaussagen.[134, 172]
Der LWMap-Service gewichtet in einem für sich abgeschlossenen Bereich. Der Betrachtungshorizont wird zum Maß der Dinge.
Es wird eine geschlossene Welt aus der Gesamtmenge der Daten gezogen. Dies hat zur Folge, dass es eine lokale geschlossene Welt gibt und den Rest, die bekannte Welt und alles Unbekannte. Sie wird durch das Retrieval vom Rest der DBPedia Page-Links abgetrennt. Relevanz ist also das Verhältnis von der abgeschlossenen Welt zur bekannten Welt, also zu den Links, die im Betrachtungshorizont liegen und dem, was in dieser Betrachtung unbekannt ist.
Numerisch wird dies durch das Verhältnis der Menge der Links im Sample zur Menge der Links in der gesamten Wikipedia dargestellt. Dieses Maß ist an den lokalen Clusterkoeffizienten[260] angelehnt. Daraus ergibt sich ein einfaches Bewertungsschema für die Wichtigkeit eines Artikels bzw. des Knotens.
Umso mehr In- und Out-Links in einem Betrachtungsraum vorhanden sind, umso wichtiger ist ein Artikel.

[17] Siehe Abschnitt 4.1.2.

Einige Rechenbeispiele hierzu:
Kontext A hat 100.000 eingehende und 100 ausgehende Links. Im vom Retrieval erstellten Subgraphen hat der Knoten 20 eingehende und 3 ausgehende Links. Damit ist nur eine sehr kleine Anzahl der Links in der Visualisierung repräsentiert.
Kontext B hat 100 eingehende Links und 30 ausgehende Links. Im vom Retrieval erstellten Graphen hat auch er 20 Eingehende und 3 ausgehende Links.
Kontext B ist also besser repräsentiert als Kontext A, da prozentual mehr seiner Links repräsentiert sind. Somit ist der gesamte Kontext wichtiger für den Knoten und der Knoten wichtiger für den Kontext.
Dies beinhaltet meistens wenig Links im Artikel und wenig Links zum Artikel. Es befindet sich damit meist am Rande des von der Wikipedia Erfassten und ist speziell. Auch wenn nur wenige Links im Betrachtungsraum auf den Artikel verweisen, wird er hoch gewichtet und damit groß dargestellt.
Artikel, die einen sehr großen In-Degree haben, wie beispielsweise die United States of America oder andere Länder, werden meistens klein dargestellt. Sie haben viele Links und über sie kann man meistens nicht sagen, dass sie etwas Außergewöhnliches darstellen. Also, auch wenn viele der in der Darstellung abgebildeten, Links zu diesem Artikel führen, ist das Visualisierte höchstens eine Randnotiz.

Richtung

Die Richtung der Links wird in gewichtete, ungerichtete Kanten überführt und als assoziatives Maß bewertet. Dabei werden auch verbundenen Knotentypen mit einbezogen. Durch diesen Schritt soll die Bedeutung für die einzelnen Links bewertet und die algorithmische Positionierung der Knoten parametrisiert werden.

Die Verbindungen zwischen Ausgangsartikeln und Kontextartikeln werden in drei Ausprägungen gewichtet.
Am stärksten wird die gegenseitige Verlinkung gewichtet, da sie wie beschrieben ein hohes Maß an Zusammenhang zeigt. Wenn Wikipedia-Artikel als eine Art Definition einer Entität gesehen werden, sind gegenseitig verlinkte Artikel sehr stark in ihrer Beschreibung voneinander abhängig.
Die Verbindung von Ausgangsknoten zu Kontextknoten wird am zweithöchsten gewichtet. Die Referenz von einer Ausgangsentität bewirkt damit in der Positionierung ein stärkeres verhältnis zu dieser speziellen Ausgangsentität.
Die schwächste Verbindung tritt bei der Verlinkung von Kontext zu Ausgangsartikel auf. Hier ist die Verbindung nicht wirklich stark und der Kontext könnte sich auf den Ausgangsknoten beziehen, ist aber für den Ausgangsknoten irrelevant. Er könnte auch als Trittbrettfahrer gesehen werden, der eigentlich marginales Wissen beschreibt, anders als der Ausgangsknoten. Daher wird die Verbindung als schwach interpretiert.
Die Verlinkungen zwischen Kontextartikeln werden in zwei Ausprägungen gewichtet.

Bei der Verbindung von Kontexten untereinander geht die Richtung nicht in die Gewich-

tung ein. Hin- und Rücklinks werden jedoch höher bewertet als die Summe der einzelnen Links, da hier auch ein besonders starker Zusammenhang angenommen werden kann, der sich auch in der Darstellung widerspiegeln soll.

5.3.5 Darstellung

Die Positionierung der Knoten und Kanten wird nach dem Retrieval vorgenommen. Diese Positionierung hat den Vorteil, dass alle Knoten eine initiale Ausgangsposition haben. Die Struktur, die aus dem Retrieval hervorgeht, wird direkt in der Positionierung der Knoten in der Visualisierung weiter verwendet.
Dabei wird eine feste Positionierung gewählt, da die daraus hervorgehende Visualisierung auch dynamisch filterbar sein soll. Die Positionierung einzelner Knoten kann also bei der Darstellung nicht verändert werden.
Die statische Positionierung hat folgende Vorteile:

Allgemeine Ortientierung
: Knoten können aufgrund ihrer Position wiedergefunden werden. Die feste Positionierung ist eine Orientierungshilfe um eine Grundorientierung zu schaffen.

Vergleichbare Ansichten
: Der Abbildungsraum ist für alle Ansichten gleich aufgeteilt.

Eine Ansicht für alle Nutzer
: Die Ansicht kann diskutiert werden, da alle Nutzer das Gleiche sehen.

Ressourcen auf dem Server
: Bei einer durchgehend dynamischen Visualisierung werden meist sehr viele Clientressourcen in Form von Rechenzeit verbraucht. Dies kann bei schwachen Client-Systemen wie Mobilgeräten problematisch sein.

Dem stehen die folgenden Nachteile entgegen:

Nicht immer optimale Darstellung
: Die Darstellung der Netze ist nicht immer optimal, in Hinblick auf Knoten- und Kantenüberlagerungen.

Server Infrastruktur
: Durch die serverseitige Berechnung müssen Rechenkapazitäten bereitgehalten werden.

Die Positionierung der Knoten erfolgt in drei Schritten. Das Vorgehen trägt der Struktur der Daten und der Laufzeit rechnung.

Positionierung der Ausgangsknoten

Da es sich bei diesem Tool um eine Karte handelt, die anhand der Ausgangsknoten erstellt wird, ist es wichtig, wie die Ausgangsknoten positioniert sind, damit die Kontextknoten möglichst überschneidungsfrei positioniert werden können. Sie sollten orientierung bieten, dem Kontext jedoch die Freiheit geben sich optimal zu verteilen.
Es werden mehrere Prämissen festgelegt:

1. Knoten, die viel miteinander zu tun haben, sind nah beieinander.
2. Knoten, die wenig miteinander zu tun haben, liegen weit auseinander und beeinflussen sich so wenig wie möglich in ihrer Ausgangsposition.
3. Es muss genug Platz zwischen zwei Knoten sein, damit die Kontexte zwischen ihnen Platz finden.
4. Ausgangsknoten, die viel miteinander zu tun haben, werden näher beieinander dargestellt.

Prämisse 1 und 3 widersprechen sich, obwohl es Argumente für beide Positionen gibt. In die Positionierung der Ausgangsknoten gehen also die Kanten von Typ 1 und das Two-Mode Netz aus Kanten vom Typ 2 und den Kontextknoten ein.
Das Two-Mode Netz der Kanten aus Typ 2 wird projiziert. Das hat eine starke Minimierung der Kanten zur Folge, was sich positiv auf die Laufzeit des Algorithmus zur Positionierung der Kanten auswirkt.
Die Kanten des Typs 1 fließen als positives, also anziehendes Gewicht mit in das Modell ein.
Nach der Positionierung der Ausgangsknoten werden sie fixiert, um ein weiteres Verschieben der Knoten zu verhindern.

Erste Positionierung des Kontexts

Durch die Positionierung der Ausgangsknoten im vorhergehenden Schritt ist eine Art Spielfeld entstanden. Dieses Spielfeld teilt den Raum zwischen den Ausgangsknoten auf. Zur weiteren Positionierung werden die Kanten des Typs 2 genommen.
Aus den ungewichteten Kanten des Typs 2 gehen mehrere mögliche Positionen hervor. Sie entsprechen den möglichen Schnittmengen zwischen den Knoten. Auf diese Weise kann ein Kontextknoten zwischen 2 bis 5 Ausgangsknoten liegen. Bei drei Ausgangsknoten lassen sich die möglichen Positionen in einem Venn-Diagramm beschreiben. Bei den Ausgangsknoten A1 bis A3 lassen sich die Kombinationen A1-A2, A2-A3, A1-A3 und A1-A2-A3 bilden. Dies ist in Abbildung 5.6 aufgezeigt.

Zweite Positionierung des Kontexts

Eine zweite Positionierung zur Verfeinerung der Position wird durch das Hinzufügen der Kanten des Typs 3 erreicht. Die Kanten des Typs 3 haben wenig Gewicht. Die Knoten lassen sich auf diese Weise differenzierter positionieren.

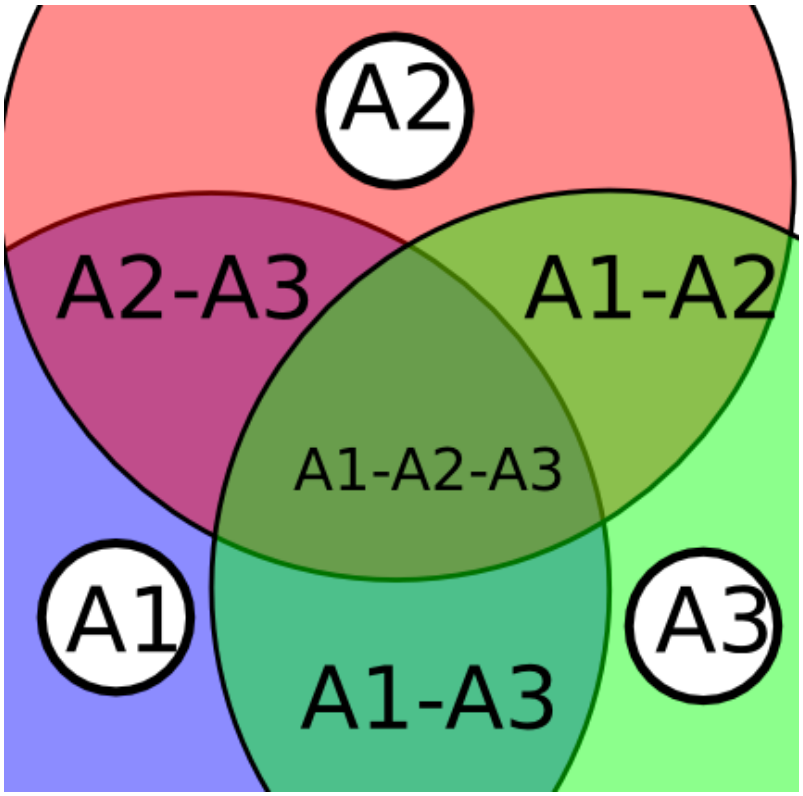


Abbildung 5.6: *Beispiel der möglichen Positionen des Kontexts anhand eines Venn-Diagramms.*

Ein Problem, das bei diesem Vorgehen auftritt, ist die Unvorhersagbarkeit der Kantenmenge des Typs 3. Wenn der Kontext sehr dicht ist und durch eine Menge von Kanten zusammengezogen wird, dann bildet sich in der Mitte der Visualisierung ein Klumpen. Das soll vermieden werden, da in diesem Fall die Knoten nicht mehr auseinandergehalten werden können und der Mittelpunkt sehr unübersichtlich wird, außerdem geht die Information die durch die Position der Kontexte vermittelt wird verloren.
Um dies zu verhindern, muss die Summe der Kantengewichte vom Typ 3 angepasst werden.Das geschieht mit der Anpassung des Gewichtsverhältnisses der Kantengewichte der Typen aneinander. Auf diese Weise lassen sich die Kräfte, die von beiden ausgehen, ausgleichen, auch wenn es sehr viel mehr Kanten des Typs 3 gibt.
Die Summe der Kantengewichte vom Typ 3 soll nicht größer werden als die Summe der Kantengewichte vom Typ 2. Die Anziehungskrägte der beiden Kantentypen sollen sich im Gleichgewicht befinden. Daher wird ein Faktor errechnet um den alle Kantengewichte des Typs 3 verringert werden.[18]

5.3.6 Nachhaltige Bereitstellung im WWW

Viele Web-Projekte aus dem akademischen Bereich verschwinden nach ihrem Projektende. Sie stellen die Funktion ein und es bleibt ein Gerippe aus HTML-Fragmenten übrig, deren

[18] Hier entsteht ein Fehler durch die endliche Genauigkeit der Fließkommazahlen. Dieser Fehler kann jedoch im Rahmen dieser Arbeit nicht weiter diskutiert werden.

dynamische Mechanismen mit der Zeit aufgehört haben zu funktionieren.
Die Gründe dafür kann man in der Projektmittelvergabe, der technischen Pflege und der Bereitstellung von nachhaltiger Infrastruktur suchen. Dabei wird bei diesen Verfahren anscheinend nicht die Zeit im Auge behalten. Durch eine Konzentration auf traditionelle Veröffentlichungen von Print-Publikationen ist der dynamische Anteil einer Arbeit im Hinblick auf seine nachhaltige Bereitstellung kaum attraktiv. Daher besteht bei komplexen Systemen die akute Gefahr, dass sie ohne Pflege nach einiger Zeit den Betrieb einstellen. Dieses Phänomen wird als Lebenszyklus von Daten[19] beschrieben. Nach dem Ablauf eines Projekts ist es eher unwahrscheinlich, dass die Pflege und die Aufrechterhaltung komplexer Systeme weiterfinanziert wird.[116]
Die nachhaltige Bereitstellung und die Reproduzierbarkeit wurden in Abschnitt 5.1 beschrieben. Im Hinblick auf Publikationen aus dem akademischen Bereich ist dabei eine nachhaltige Bereitstellung von dynamischen Visualisierungen besonders interessant, da so Argumente untermauert und kritisiert werden können.
Um dem Phänomen des Datenverfalls entgegen zu wirken, gibt es Möglichkeiten die Komponenten einer Applikation in verschiedene Kategorien, die mehr oder weniger statisch bereitgestellt werden können, aufzuteilen. Dabei fallen je nachdem Metainformationen über den Ablauf zur Reproduzierbarkeit an. Diese Eigenschaften im Blick soll im Weiteren eine Strategie entworfen werden, wie Ergebnisse nachhaltig darstellbar sind. Dabei sollen komplexe, pflegebedürftige Systeme von den statischen Systemen entkoppelt werden.

Statische und dynamische Komponenten

Es gibt rechenintensive spezialisierte serverseitige Prozesse. Dazu gehören die Bereitstellung im Internet, der Betrieb und die Pflege von Datenbanken, das Betreiben der Verarbeitungslogik und komplexen Algorithmen die von Klienten, wie Mobilgeräten nicht bewältigt werden können.
Zu diesen klar serverseitigen Prozessen kommt das Interface, welches zwangsweise auf den Endgeräten bereitgestellt werden muss. Der Client konsumiert die Daten und stellt diese auf einem Display für den Nutzer dar. Je nach Form der Daten muss der Client die Visualisierung selbst berechnen oder einfach nur darstellen. Dabei können auf dem Server Vorberechnungen für die Darstellungen angestellt werden. Die Menge dieser Vorberechnungen und deren Sinnhaftigkeit müssen Abgewägt werden.
Ein Web-Browser kann im simpelsten Fall eine Pixelgrafik darstellen, die vom Server berechnet wurde, das wäre eine Möglichkeit der statischen Abbildung der Daten. Dabei ist die Darstellung als Vektorgrafik jedoch flexibler verbraucht weniger Overhead. Auch lassen sich so einzelne Elemente klar adressieren, was für eine interaktive Visualisierung notwendig ist.
Wenn, wie im Falle von JavaScript, eine dazugehörige Verarbeitungs- und Manipulationslogik hinzukommt, ist eine vollständige Anwendung auf Seiten des Clients realisierbar. Dabei muss nach dem Laden der Daten nicht mehr mit einem Server kommuniziert wer-

[19] Engl.: data lifecyle

den, da der Content auch Lokal gespeichert und abgerufen werden kann.
Hier dienen HTML basierte Web-Seiten als Client Interface. Der Web-Browser und die Dateistandards erfüllen dabei den Zweck, die Programmlogik plattformunabhängig zu halten. Die Reproduzierbarkeit wird dabei in einfacher Form aufrechterhalten, da JavaScript Dateien quell-offen sind[20]. Durch die offene Standardisierung und Implementierung sowie die Verbreitung kann angenommen werden, dass es lange zeit Engines geben wird, die JavaScript mit HTML verarbeiten können werden.

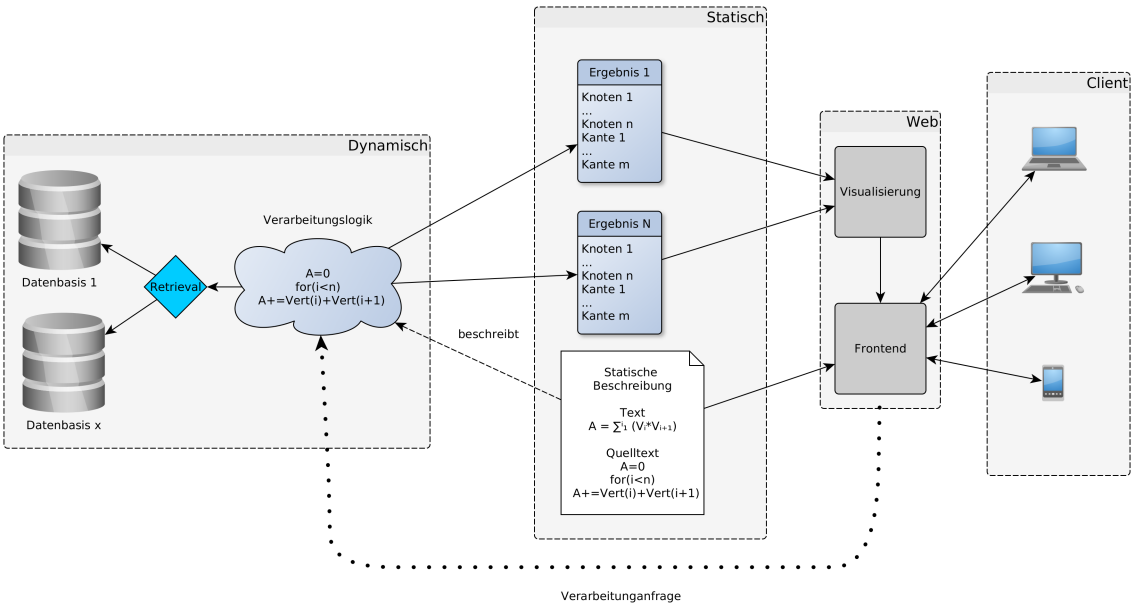


Abbildung 5.7: *Die für den LWMap Service implementierte teilstatische Architektur.*

Eine Beschränkung dieses Ansatzes ist die Menge der darstellbaren Knoten und Kanten. Dies hat mit dem hohen Abstraktionsgrad der Darstellung zu tun. Die Abstraktion erzeugt durch viele menschenlesbare Technologien einen großen Overhead an Daten und Berechnungen. Bei mehr nativen Anwendungen wie kompiliertem Code ist eine solche Verarbeitung sehr viel effizienter[21] und auf gleicher Hardware ist die Darstellung von mehr Knoten und Kanten möglich. Diese Performanceprobleme werden hier aber nur teilweise kritisch, da Netzwerkdarstellungen mit steigender Anzahl von Knoten und Kanten generell schwerer zu verstehen sind.
An dieser Stelle lassen sich zwei Komponenten entkoppeln:
1. das serverseitige Backend, das Pflege und Administration bedarf.
2. die Informationen, wie sie an den Client als statische Daten übertragen werden.
Das Retrieval und die Datenhaltung der Rohdaten können auf diese Weise ausfallen, oh-

[20] JavaScript ist eine Scriptsprache, der Quellcode wird vom Browser immer wieder neu interpretiert. Quellcode kann jedoch optimiert werden, so dass Maschinen ihn interpretieren können, Menschen jedoch nicht mehr lesen. Dieses Vorgehen wird zur Optimierung verwendet. Kommentare und Variablennamen und Zeilenumbrüche werden auf ein Minimum reduziert. Auf diese Weise wird der Netzwerkverkehr einer Website zu minimiert.

[21] Hier ist beispielsweise Flash zu nennen. Flash liegt nur als Binärcode vor und auch in offenen Implementationen ist sie kein Teil des W3C Standards. Sie wird meist durch proprietäre Plug-ins in Browsern eingebunden. Diese Technologie wird beispielsweise beim RelFinder verwendet.

ne die Bereitstellung von gespeicherten Suchergebnissen zu beinträchtigen. Der statisch erstellte Content kann durch einen simplen Webserver ausgelieferten werden. Ein HTML-basiertes Interface mit statischem Content muss nicht online bereitgestellt werden. Ein Nutzer kann die Ergebnisse als Abbild der vorliegenden Homepage kopieren und diese lokal verwenden.
Ein Webserver wird als Standardwerkzeug gesehen, das langlebiger ist als beispielsweise experimentelle Datenbanken und der Etat von kurzweilig finanzierten Projekten. Das macht diese Lösung teilnachhaltig. Die rechenintensiven Technologien im Hintergrund sind dabei vom bestehendem Content entkoppelt.
Eine Website, die nur Bestehendes darstellen und bereitstellen kann, könnte als Diorama bezeichnet werden. Es ist eine Momentaufnahme, in der mehrere Perspektiven eingenommen werden können. Die Analogie der Perspektive wird auf Filter und den Visualisierungsalgorithmus bezogen. Eine Ansicht ist ein Argument, ein herausgestellter Aspekt.
Wie sich ein digitales oder analoges Foto zur Welt verhält, so bezieht sich eine konkrete Visualisierung auf eine Datenlage, die einem Zustand zu einem Zeitpunkt entspricht. Die Perspektive des Bilds entspricht dabei der konkreten Visualisierung und dem, was die Visualisierung versucht sichtbar zu machen. Diese Perspektive soll jedoch auch argumentativ nachhaltig genutzt werden können. Hier bietet Linked Data eine Möglichkeit, auch auf der Datenseite den Menschen persönliche gewählte Ansichten auf der Grundlage des gleichen, langzeitverfügbaren Weltwissen zu erstellen.

Backendtechnologien und Prämissen

Im Backend des LWMap Service kann die Technologie in mehrere Teilschritte aufgeteilt werden.
Es gibt die Datenhaltung, das Retrieval und die Visualisierung, dazu kommt ein anonymisierter Ansatz zum User-Tracking.
In der Datenhaltung wird der Fuseki-Server[9] verwendet, er kann die von DBpedia bereitgestellten Dumps direkt einlesen und ist simpel zu konfigurieren. Das Datenretrieval erfolgt über eine SPARQL-Endpoint. Dieser wird vom Fuseki Server bereitgestellt.
Das Backend, das diesen SPARQL-Endpoint abfragt, besteht aus einem Java Programm, es enthält die Logik des Retrievals, der Datenkomposition und der Visualisierung mithilfe des Gephi Toolkits[22]. Diese Kernkomponente kann auch durch vergleichbare Technologie ersetzt werden. Die Arbeit des Backends ist in 5.3.3 bis 5.3.5 beschrieben.
Die Anfragen und das User-Tracking werden von einem Backend in node.js[186] bereitgestellt, das massiv parallele Anfragen verarbeiten kann und serverseitig ausgeführtem JavaScript entspricht.
Alle technologien sind unter einer freien Lizenz verfügbar.[23]

[22]Die Funktionalität des Programms Gephi[19] als Java API.

[23]Java wird über die offene OpenJDK Implementierung verwendet.

Frontendtechnologien und Prämissen

Die Wahl der Technologien erfolgte in diesem Fall nach dem Paradigma der Nachhaltigkeit. Es wurden daher Web-Technologien aus dem HTML5-Standard gewählt. Durch ihre offene Dokumentation und weite Verbreitung sind diese Standards wahrscheinlich auch in Zukunft noch interpretierbar. Die Konsequenz daraus ist, dass die informationstechnischen Objekte, die aus dem Projekt entstehen, möglichst ohne äußere Pflege und mit einem Minimum an Infrastruktur auskommen.
Das heißt, sie sind funktionstüchtig ohne dass eine verwaltungsintensive serverseitige Technologie dahinter stehen muss, also Datenbanken und andere serverseitig betriebene Spezialsoftware.
Die Unabhängigkeit und weitläufige Verwertbarkeit wird durch den Web-Browser gewährleistet. Browser sind auf fast allen Computern lauffähig und auf fast allen Client-Rechnern standardmäßig installiert. Daher bieten sie die perfekte technologische Grundlage zur nachhaltigen Arbeit mit Daten.
Die dazugehörige Technologie besteht aus HTML und JavaScript. Diese Standards werden im Browser ausgeführt und nicht auf dem Server selbst. Es müssen also nur Dateien bereitgestellt werden, um die Ergebnisse anzuzeigen. Des Weiteren ist keine kommerzielle Software nötig. Die Software kann also immer interpretiert und korrigiert werden, da sie und die Dokumentation öffentlich verfügbar sind. Die Entwicklung von JavaScript und die Entwicklung der Endgeräte lässt auf eine Tendenz zur weiteren Nutzung von JavaScript schließen. HTML5-Elemente wie WebGL, die eine clientseitige hardwarenahe Visualisierungsebene ermöglichen, zeigen dabei neue Möglichkeiten auf.
Im Bereich der Webvisualisierung kann man auch Initiativen zur intelligenten Weiterentwicklung erkennen. Es gibt Anstrengungen die Ausführung von JavaScript performanter zu machen beispielsweise durch Standard konforme Simplifizierung wie ASMjs.[10] ASMjs basiert auf einer Syntax-Untermenge von JavaScript, das durch seine Einschränkungen von Spezialinterpretern performanter ausgeführt werden kann. Es kann aber auch von einfachen JavaScript Interpretern ohne eine Spezialinterpretation verarbeitet werden. Eine weitere Möglichkeit ist es JavaScript-Code mit Emscripten aus anderen Sprachen zu erzeugen.[271] Eindrucksvoll belegt dies die Retrodigitalisierung von über 2000 DOS-Computerspielen durch das Internet Archive, die auf diesen Technologien beruht.[235] Auf diese Weise wurden alte Spiele direkt im Browser spielbar und sind damit weiterhin erfahrbar. Ausserdem gibt es weiterhin JavaScript basierte serverseitige Entwicklungen. Neben node.js ist hier Coffee-Script zu nennen.
Die gefestigte Position im Web-Umfeld und die konzeptionelle Quelloffenheit von nicht serverseitigem JavaScript macht es zum passenden Mittel für die angestrebte Nachhaltigkeit. Die Ergebnisse der Verarbeitung im Backend werden im Graph Exchange XML Format, kurz Gexf[104], bereitgestellt, da das im Hintergrund arbeitende Framework auf Gephi beruht. Gexf ist wie die meisten XML basierten Standards offen dokumentiert.
Zur Darstellung wird Sigma.js[5] genutzt. Diese Visualisierungsbibliothek kann Gexf Dateien nativ parsen und darstellen. Sigma.js zeichnet sich desweiteren durch seine einfache Anpassbarkeit aus. Viele der Grundfunktionen sind standardmäßig implementiert, wie

das Zoomen und die Interaktivität der Grafik. Einzelne Funktionselemente können aber einfach angehängt und durch selbst geschriebene Elemente ersetzt werden. Dies betrifft beispielsweise Renderer[24] für Knoten, Kanten, Text u.s.w. Technologisch basiert Sigma.js auf JavaScript und dem Canvas Element aus dem HTML 5 Standard und bedient damit die erwähnten Webstandards. Es ist jedoch auch möglich, WebGL als alternative Visualisierungstechnik zu nutzen. Das setzt jedoch voraus, dass alle Renderer auch in WebGL implementiert sind. Auf WebGL Renderer wurde daher verzichtet.
Die Auffindbarkeit und Wiederverwendbarkeit sowie das Teilen von bestimmten Ansichten auf die Daten wird im statischen Frontend vorgenommen. Eine erstellte Visualisierung wird als Datei hinterlegt. Ein Hash, der sich aus den wichtigen Informationen für die Visualisierung errechnet, dient zur Benennung der Datei. Aus Suchanfragen kann also immer der Dateiname hergeleitet werden.
Der Hash wird aus den URIs der Ausgangs-Entitäten in sortierter Reihenfolge erzeugt. Auf diese Weise kann die Reihenfolge der Eingabe nicht das Ergebnis des Hashings beeinflussen. Auch werden die Datenversion und die Version der Backendsoftware zum Generieren des Hashes einbezogen. Auf diese Weise verhindert der Hash, das doppelte Visualisierungen erzeugt werden. Wird eine identische Anfrage bei gleicher serverseitiger Verarbeitungslogik ein zweites Mal abgeschickt, wird die vorhandene Datei angezeigt. Beim Öffnen einer Ansicht wird der Hash an die URL angehängt.
Um die Ansichten und Frontend-Filter zu teilen, wird am gleichen URL in eine serialisierte Form der Filterparameter angehängt. Dieses Vorgehen ist nur bis zu einer begrenzen Länge der parameter Praktikabel[25], so werden beim Weitergeben keine statischen Daten erzeugt und auf dem Server hinterlegt. Auf diese Weise ist das Ergebnis immer mit der gleichen URL aufrufbar. Es entsteht eine Art URI.

Interaktivität der Visualisierungen

Eine Zusatzinformation, die dem lokalen Wiki-Graphen hinzugefügt wurde, ist die Klasse von Wikipedia Artikeln, wie sie von DBpedia bereitgestellt wird. Die Artikel können so nach ihren Typen gefiltert werden, also nach Personen, Orten, Firmen etc. . Die Ontologie, die zugrunde liegt, ist aus den Infoboxen der Wikipedia extrahiert und somit sind auch mehrfach Kategorisierungen möglich. Diese Filter zeigen auch auf, welche Artikel nicht typisiert sind.

Hidden Context

Ein Feature, das zum Filtern oder Hervorheben genutzt werden kann, ist der Hidden Context. Dabei handelt es sich um Knoten, die beim einfachen Browsen nicht erfasst werden können. Es sind Artikel, die nur auf die Ausgangsartikel verlinken, jedoch von keinem

[24] Renderer sind Programmteile die sich mit dem Zeichnen von abstrakten Informationen in Pixelgrafiken, beispielsweise für die Bildschirmdarstellung, beschäftigen.

[25] Auch wenn diese nicht einfach zu benennen sind, kann davon ausgegangen werden, dass Browser nur eine begrenzte Maximallänge von URLs beim Aufruf zulassen.

der Ausgangsartikel referenziert werden. Diese versteckten Verbindungen können gerade für Experten aus dem geisteswissenschaftlichen Bereich interessant sein wie von Corey Angeführt.[60] Als Hidden Context werden weniger bekannte oder kleine Artikel auftauchen. Durch ihre passive Verlinkung fallen sie beim Browsen aus dem Betrachtungsraum des Nutzers heraus. Auf diese Weise können Artikel gefunden werden, die wahrscheinlich in Suchmaschinen Abfragen auftauchen würden, aber nicht zwangsweise im Aufmerksamkeitshorizont der Wikipedianer zu einem Thema liegen, da sie von den Ausgangsartikeln nicht verlinkt werden.

5.4 LWMap Auswertung

Der LWMap Service wurde Online gestellt und seit dem 24.März 2014 unter
`http://lwmap.uni-koeln.de/`
betrieben.

5.4.1 Beispiele im Vergleich

Der LWMap Service hat einige Beispiele bereitgehalten. Anhand dieser Beispiele wurde auch die Parametrisierung des Visualisierungsalgorithmus vorgenommen. Das erste dieser Beispiele besteht aus den Wikipedia-Artikeln “1920s”, “Swing_music”, “Fashion”, “Gangster”. Dieses Beispiel ist gewählt worden, um das Potential einer assotiativen Beschreibung einer Suche zu Formulieren. Zusammengefasst wird das Thema des “film noir” oder der Gangstergeschichte aus der amerikanischen Prohibitionszeit dargestellt. Wie beschrieben ergibt diese Aufstellung über die typisierten Links des RelFinders, der in der Demoversion die typisierten DBpedialinks nutzt, kein Ergebnis.
Die Einzelansicht im LWMap Service ergibt einen von der Knoten- und Kantenanzahl recht übersichtlichen Graphen wie in Abbildung 5.8 zu sehen ist. Das Beispiel ist durch seine herausgehobene Stellung sehr oft angeklickt worden, daher soll es im Weiteren als exemplarisches Beispiel für die Beschreibung der Funktion dienen.
Aus der Vollansicht sticht ein Artikel selbst ohne Zoom heraus, dies ist der Artikel zum Roman “Lady in the Mourge”[154] von Jonathan Latimer, einer Gangstergeschichte aus dem Jahr 1938.
Durch Zoomen können Ausschnitte genauer betrachtet werden und es erscheinen auch die Bezeichnungen der einzelnen Knoten, wie in Abbildung 5.9a gezeigt wird.
Um diese Visualisierung und das Netz herunter zu laden, gibt es einen Dialog, der neben der Download-Funktion auch die Möglichkeit gibt, die angezeigte Visualisierung mit den jeweiligen Filtern per Link zu teilen. In Abbildung 5.9 werden Möglichkeiten gezeigt wie einzelne Klassen von Knoten hervorgehoben werden können. Dabei ist in Abbildung 5.10a gefiltert, welche Artikel nicht klassifiziert sind, es werden nur die Unklassifizierten hervorgehoben. Auffällig ist, dass viele Artikel in der Visualisierung unklassifiziert sind.

In Abbildung 5.10c werden Personen und teilweise auch Institutionen hervorgehoben.[26] Alle hier gezeigten Fokussierungen können auch als Filter angewendet werden, alle Knoten die nicht in den Filter fallen werden dann aus der Darstellung ausgeblendet.

5.4.2 Vergleich RelFinder

Die Relationen, die bei Wikipedia als einfache Links hinterlegt sind, bilden ein breiteres Spektrum ab. Da die Datengrundlagen und die Intentionen unterschiedlich sind, kann kein Vergleich im Sinne einer "besseren" oder "schlechteren" Software angestellt werden. Es werden also eher verschiedene Ansichten einer Datenquelle geliefert, die ähnliche Interfacekonzepte verwenden. Viele der weicheren Begriffe werden beim RelFinder nicht gefunden.
Der RelFinder und die in der Demo hinterlegten Daten liefern bei der Darstellung des ersten Beispiels des LWMap Service, wie im vorheriegen Abschnitt gezeigt wurde, ein leeres Ergebnis. Der Zusammenhang, wie er in diesem Beispiel beleuchtet wird, ist dabei weniger klar durch vorliegende semantische Relationen zu beschreiben.
Das Beispiel der Kubakrise schafft einen Zusammenhang, wie in Abbildung 5.11 zu sehen. Dabei wird jedoch die Zeit, der Artikel zu den "1960s", nicht mit einbezogen. Das Beispiel ist im LWMap Service sehr dicht, beim RelFinder wird die größere Informationsdichte sichtbar. In der Beispielliste des RelFinders werden die Entitäten "Fiat", "Maserati" und "Ferrari" in Verhältnis gesetzt. Diese Darstellung ist im LWMap Service reichhaltiger als im RelFinder, wie in Abbildung 5.12 zu sehen ist. Das Vorgehen im LWMap Service filterte automatisch, um das Ergebnis nicht zu überfrachten. Es wurden die Relationen zwischen Ferarri und Maserati sowie zwischen Ferrari und Fiat gefiltert.[27] Im Generellen kann gesagt werden, dass die Verknüpfungen in den Daten des RelFinders nicht so eng sind wie im LWMap Service. Das hat mit der Datengrundlage zu tun. Dabei ist das Vorgehen an die Menge der Daten, die einer Visualisierung zugrunde liegen, gebunden.
Aufgrund der Datengrundlage sind die Verknüpfungen in den Daten des RelFinder nicht so dicht wie im LWMap-Service, dieser stellt eine Übersicht über untypisierte Links her. Hier kann eine Analogie mit Precision und Recall gezogen werden. Der RelFinder hat eine höhere Precision und dies nicht nur im Sinne des Information Retrival, sondern auch im semantischen Sinne. Der LWMap Service liefert eine größere Menge an Ergebnissknoten zurück.

[26] Die zu diesem Beispiel gehörende URL `http://lwmap.uni-koeln.de/display.html?existingFile=20sSwingandCrime_en&FilterString=%5B%7B%22value%22%3A%22AgentPerson%22%2C%22field%22%3A%22classes%22%2C%22distinct%22%3Afalse%7D%2C%7B%22value%22%3A%22ArtistPerson%22%2C%22field%22%3A%22classes%22%2C%22distinct%22%3Afalse%7D%2C%7B%22value%22%3A%22Agent%22%2C%22field%22%3A%22classes%22%2C%22distinct%22%3Afalse%7D%2C%7B%22value%22%3A%22MusicalArtist%22%2C%22field%22%3A%22classes%22%2C%22distinct%22%3Afalse%7D%5D&FilterOptions=%7B%22exclusive%22%3Afalse%2C%22highlight%22%3Atrue%7D`

[27] Details: `http://lwmap.uni-koeln.de/display.html?existingFile=2b37f3389241f64ef60358d22ba5e653`

Abbildung 5.8: *Vollansicht des Hauptbeispiels*

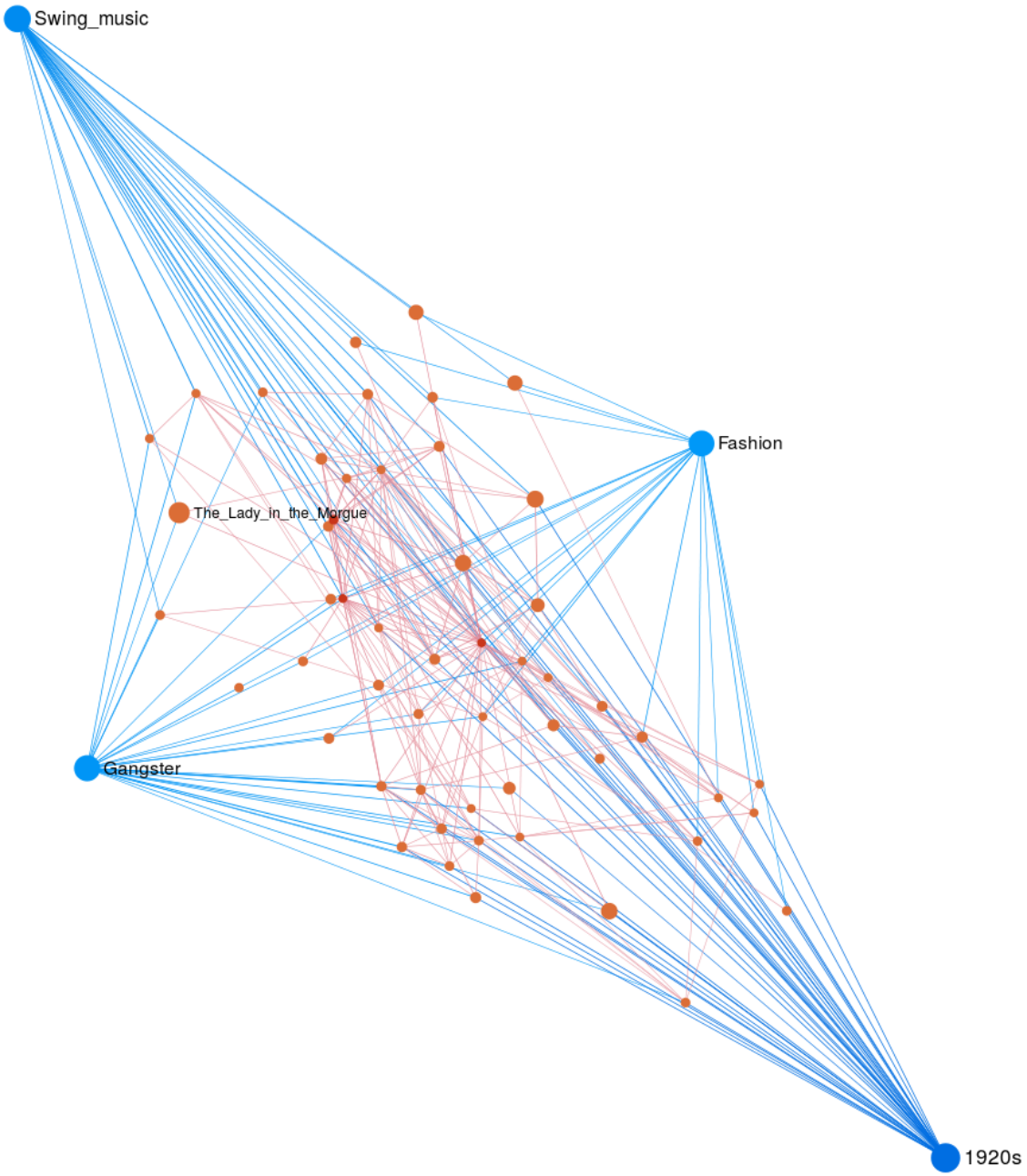
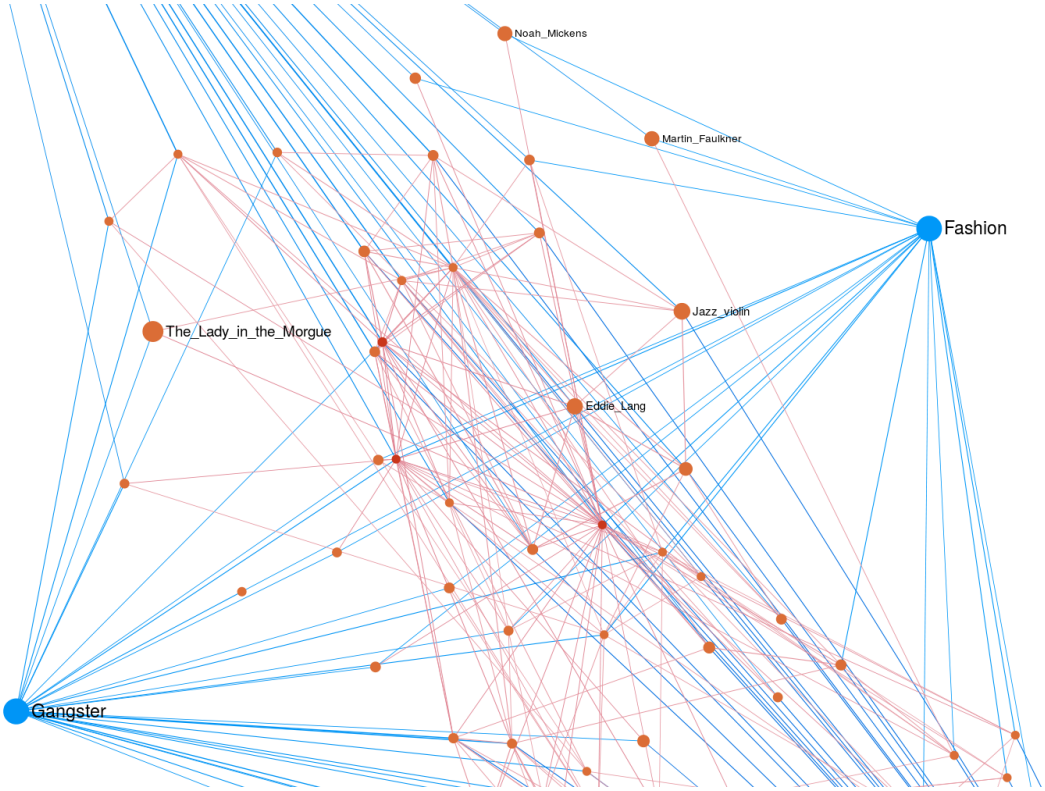
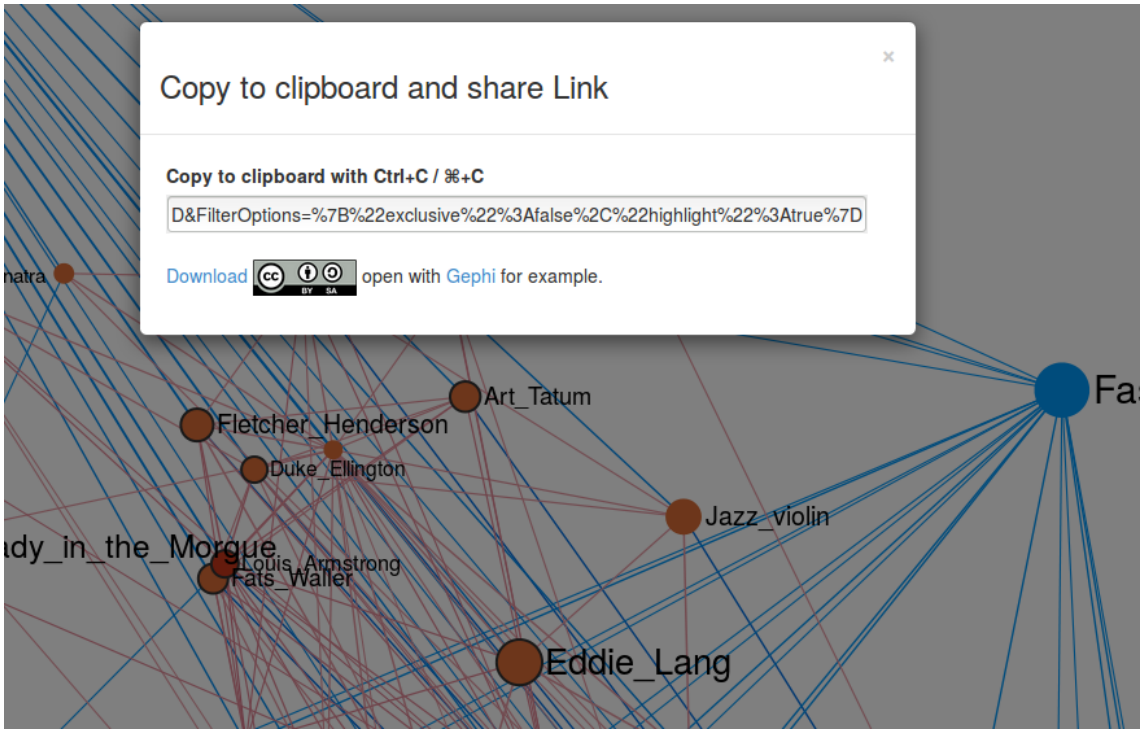


Abbildung 5.9: *Ansicht des Hauptbeispiels*

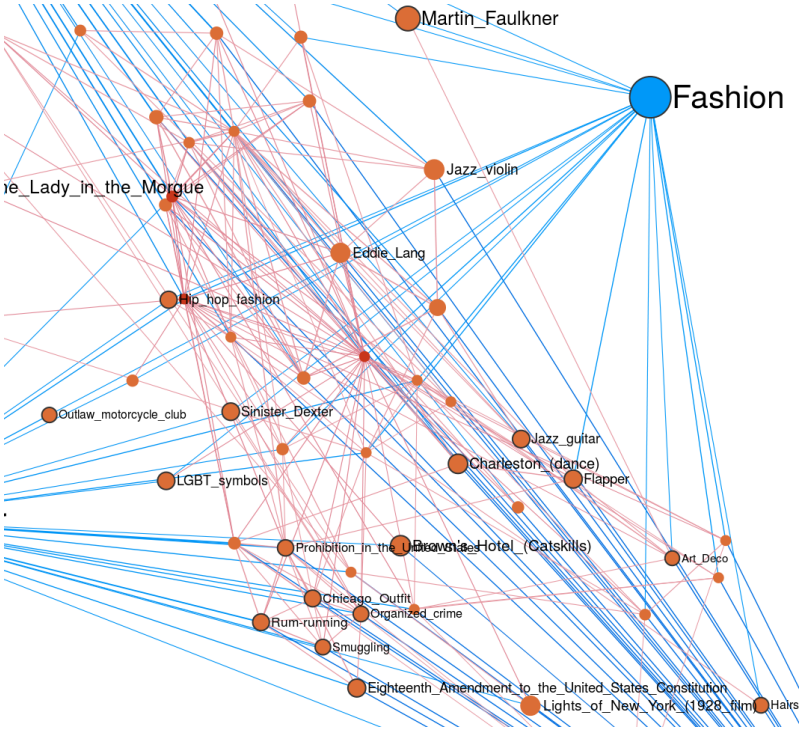


(a) Zoomansicht des Hauptbeispiels

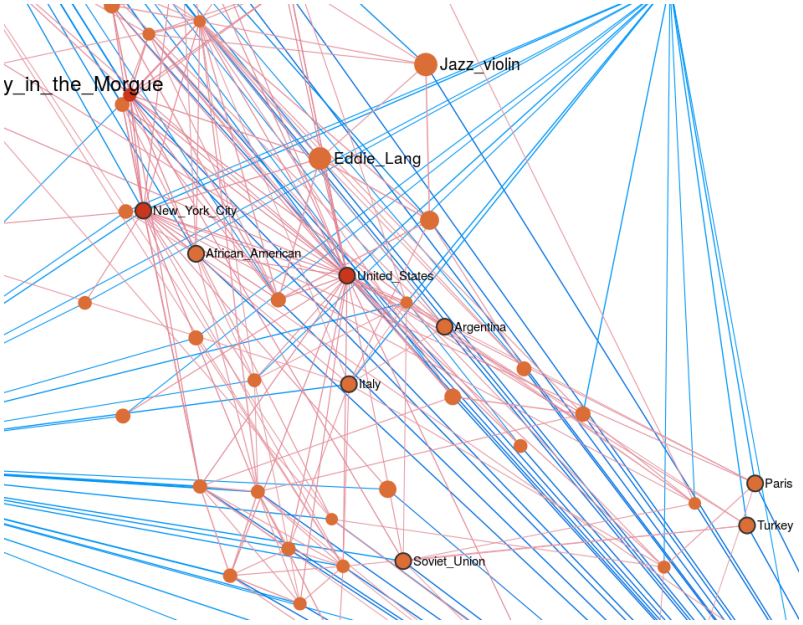


(b) "Teilen" Funktion

Abbildung 5.10: *Filter Beispiele*



(a) *Hervorhebung der Artikel ohne Klasse*



(b) *Hervorhebung der Orte beschreibenden Artikel*

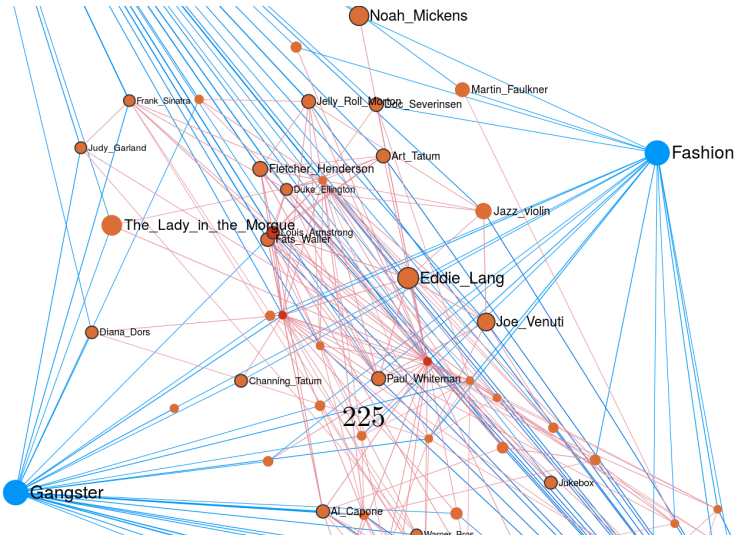
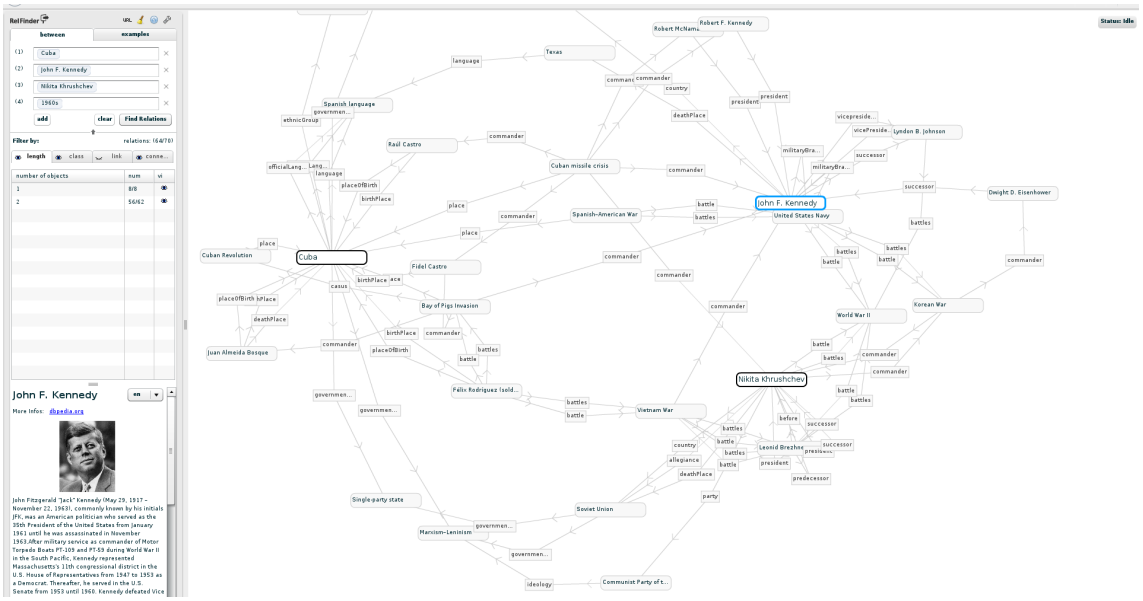
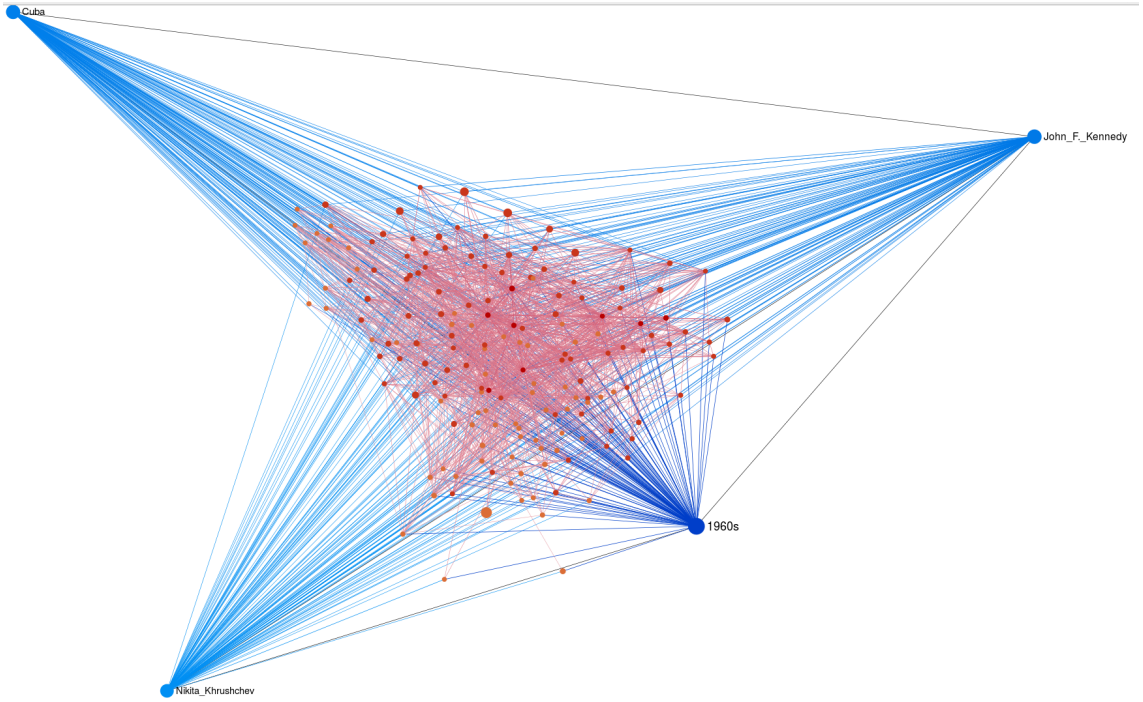


Abbildung 5.11: *Beispiel mit Cuba, John F. Kennedy, Nikita Khrushchev und 1960s*

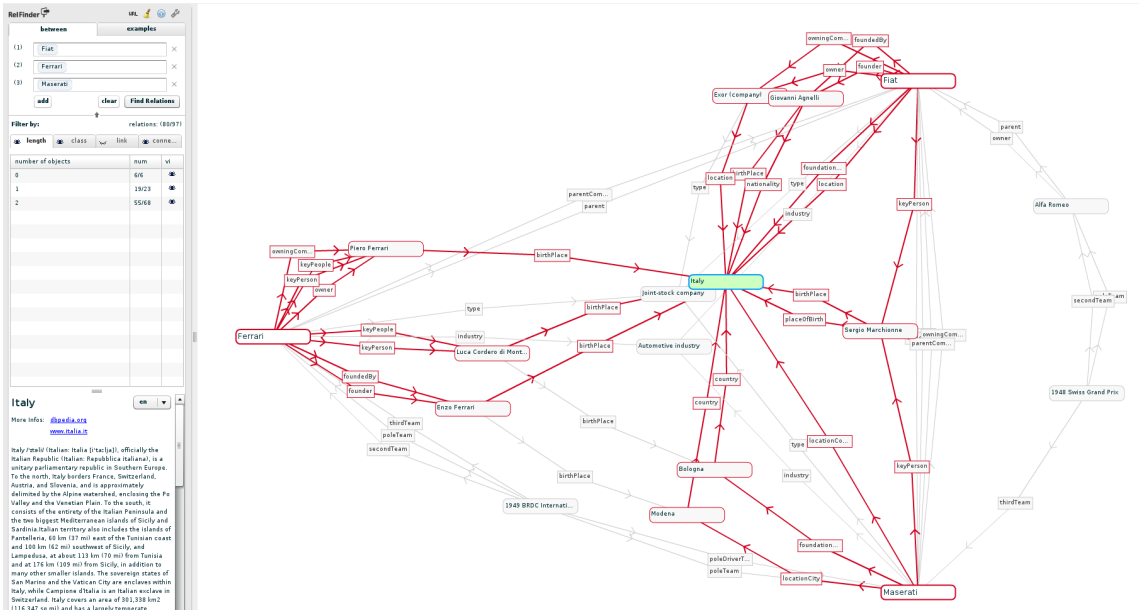


(a) *RelFinder*

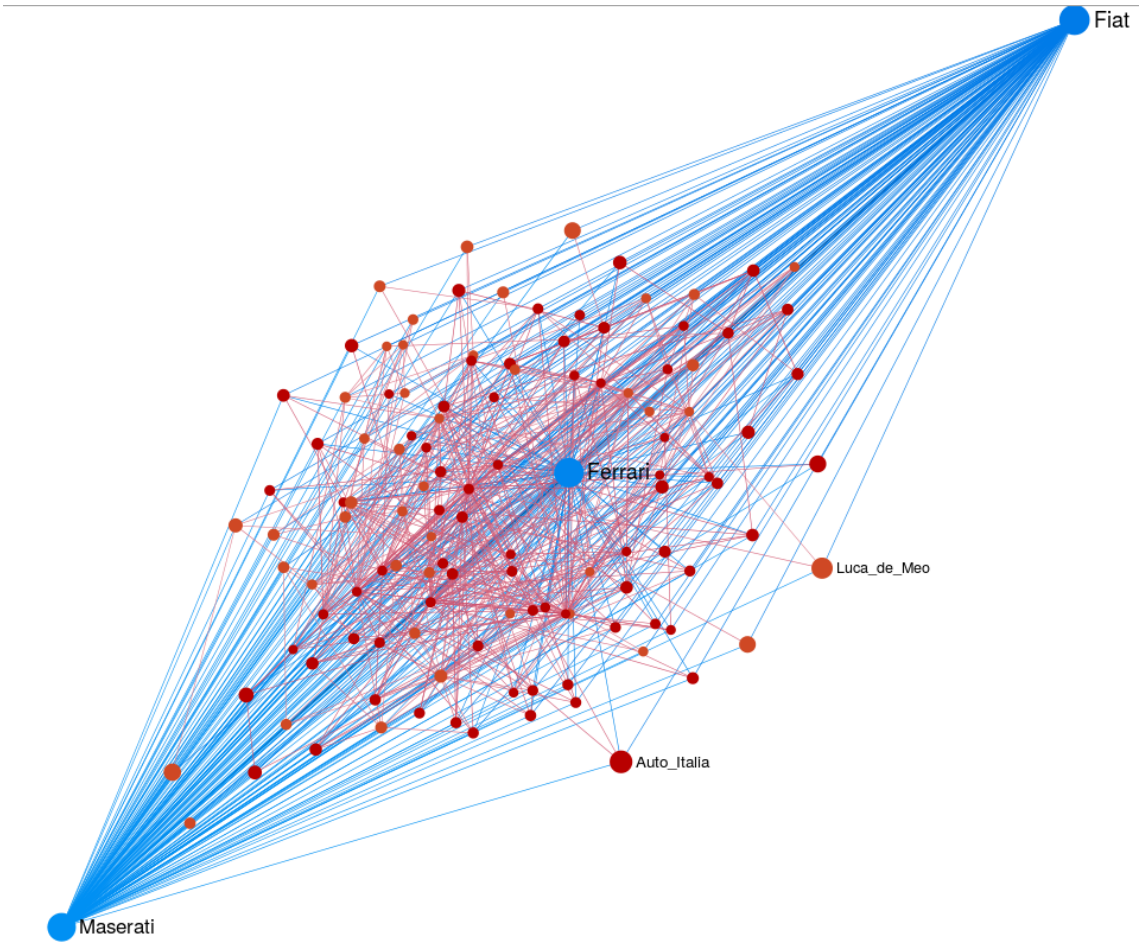


(b) *LWMap Service*

Abbildung 5.12: *Beispiel mit Fiat, Maserati und Ferrari*



(a) *Relfinder*



(b) *LWMap Service*

5.4.3 Auswertung

Der LWMap Service ist zu dem Zeitpunkt, zu dem dies geschrieben wird, noch in Betrieb. Die folgende Auswertung basiert auf Daten aus dem Zeitraum vom 24.März 2014 bis zum 31.Dezember 2014. Die folgenden Auswertungen sind auf der Grundlage des beobachteten Nutzerverhaltens entstanden. Dabei handelt es sich um die erstellten Visualisierungen, das Abrufen von Visualisierungen, das Anklicken von Knoten und das Filtern von Werten.
In besagtem Zeitraum wurden von den Nutzern 471 Visualisierungen erstellt, davon nutzen 445 die Links der englischen Wikipedia und nur 26 die Links der deutschen Wikipedia. Von diesen Visualisierungen beruhten 196 Visualisierungen auf 3 Ausgangsknoten, 342 auf 4 Ausgangsknoten und 129 auf 5 Ausgangsknoten.
In den Visualisierungen wurden 717 Mal Filter genutzt. Der am häufigsten verwendete Filter zeigte den Hidden Context einer Anfrage. Das kann teilweise dadurch erklärt werden, dass dieser Filter an prominentester Stelle im Layout zu finden war und auf der Startseite direkt angekündigt wurde.

Auswahl und Zusammenhang

In den Visualisierungen des LWMap Service wurde das Nutzerverhalten verfolgt. Die Nutzungsanalyse bezieht sich im Folgenden darauf, welche Visualisierungen erstellt wurden und mit welchen Knoten in der Visualisierung interagiert wurde.
Bei dieser Auswertung wird ein qualitativer Ansatz verfolgt, um die gesammelte Datenmenge sichtbar zu machen. Die Darstellung von Zusammenhang, Abbildung 5.13 und 5.14, zeigt, wie die im Betrachtungszeitraum erstellten Visualisierungen über ihre Ausgangsknoten zusammenhängen. Es bildet sich ein bipartites Netz von Visualisierungen, verbunden über die Ausgangsartikel, denen sie zugrunde liegen.
In der Darstellung werden dafür zwei Arten von Knoten unterschieden. Erstens die Ausgangsartikelm zweitens die Visualisierungen selbst.
Die Ausgangsartikel sind farbig zwischen Grün und Blau aufgeteilt und haben eine Beschriftung. Die Farbe gibt dabei an, ob es sich um Blätter handelt. Ausgangsartikel, die nur einmal verwendet wurden, sind blau gefärbt. Die grünen Knoten sind Ausgangsartikel, die in mehreren Visualisierungen verwendet wurden.
Die Visualisierungen sind über einen Gradienten von Beige über Orange nach Braun eingefärbt, dies gibt den Erstellungszeitpunkts der Visualisierung an. Weiß ist früh und braun ist spät im Untersuchungszeitraum. Die Zeit wurde nicht absolut gewertet sondern als Reihenfolge. Die Farben bilden die Reihenfolge der Erstellung ab und nicht die absolute Zeit. Dieses Vorgehen hilft, die Farben besser zu differenzieren, ohne direkt die Daten der Nutzer zu verwenden.
In dieser Visualisierung kann abgelesen werden, welche inhaltlichen Überschneidungen Visualisierungen haben und ob diese zeitnah erstellt wurden. Dadurch wird sichtbar, was zufällig zusammenfällt und was zeitlich nah beieinanderliegt. Die Artikel, die sich um Visualisierungen mit gleicher Farbe anordnen, sind daher wahrscheinlich Variationen von

Suchen durch ein und dieselbe Person[28] oder eine etwaige Gruppe von Personen an verschiedenen Endgeräten. Die Gruppe könnten die Visualisierungen beispielsweise direkt oder über eine Plattform wie ein Soziales Netzwerk, ein Forum oder ähnliches ausgetauscht haben.

In Abbildung 5.13 ist eine Kette von Knoten entstanden. Diese setzt sich aus verschiedenen Aufrufen zu verschiedenen Zeitpunkten zusammen. Durch die verschiedenen Farben der Visualisierungsknoten sind die Artikel, die zufällig zu verschiedenen Zeitpunkten gewählt wurden, ausmachbar.

Die Artikel “Brain” und “Mental health” (unterer Teil) sowie “Vietnam” und “Poetry” (oberer Teil) wurden wahrscheinlich unabhängig voneinander gewählt. Die Visualisierungsknoten, die sie umgeben, sind verschiedenfarbig und daher zeitlich unabhängig voneinander.

Anders sieht es für die Knoten “Science”(unten) “Anime”(Mitte) und “San José”(oben) aus. Die Visualisierungen, die sie umgeben, sind gleichfarbig, daher wurden sie zum gleichen Zeitpunkt wahrscheinlich von der gleichen Person oder Gruppe öfters genutzt.

Die beschriebene Kettenkomponente kann projiziert werden. Das ermöglicht eine Darstellung auf kleinerem Raum, da weniger Knoten abgebildet werden müssen. Dadurch können die assoziativen Pfade besser verfolgt werden. Die Komponente ohne Visualisierungen ist in Abbildung 5.15 zu sehen. Da durch die Projektion nur noch eine Art von Knoten vorhanden ist, lassen sich die Farben und die Größe neu codieren. Das lässt einen besseren Fokus auf die benennbaren Knoten zu. In dieser Abbildung wird dabei ein besonderer Wert auf die Unterscheidung von Zentralität und Grad gelegt, da so die transitive Wertung gegen die lokale Wertung vorgenommen werden kann. Die Farbintensität entspricht dabei der lokalen Wertung, dem Grad. Die Größe entspricht der transitiven Wertung, der Betweenness Centrality. Die vielfarbige Abbildung war komplexer, daher schwerer zu lesen. bei dieser Graphik kann abgelesen werden welche Knoten Zusammenhang erzeugt haben. Dabei ist jedoch nicht mehr zu sehen, ob dieser Zusammenhang zu ähnlichen Zeiten entstand ist.

[28]Hier könnte auch mit Cookies gearbeitet werden, um einzelne Nutzer zu identifizieren. Bei der Verwendung verschiedener Web-Browser, Endgeräte oder dem Löschen der Cookies im Browser ist dies jedoch nicht eindeutig.

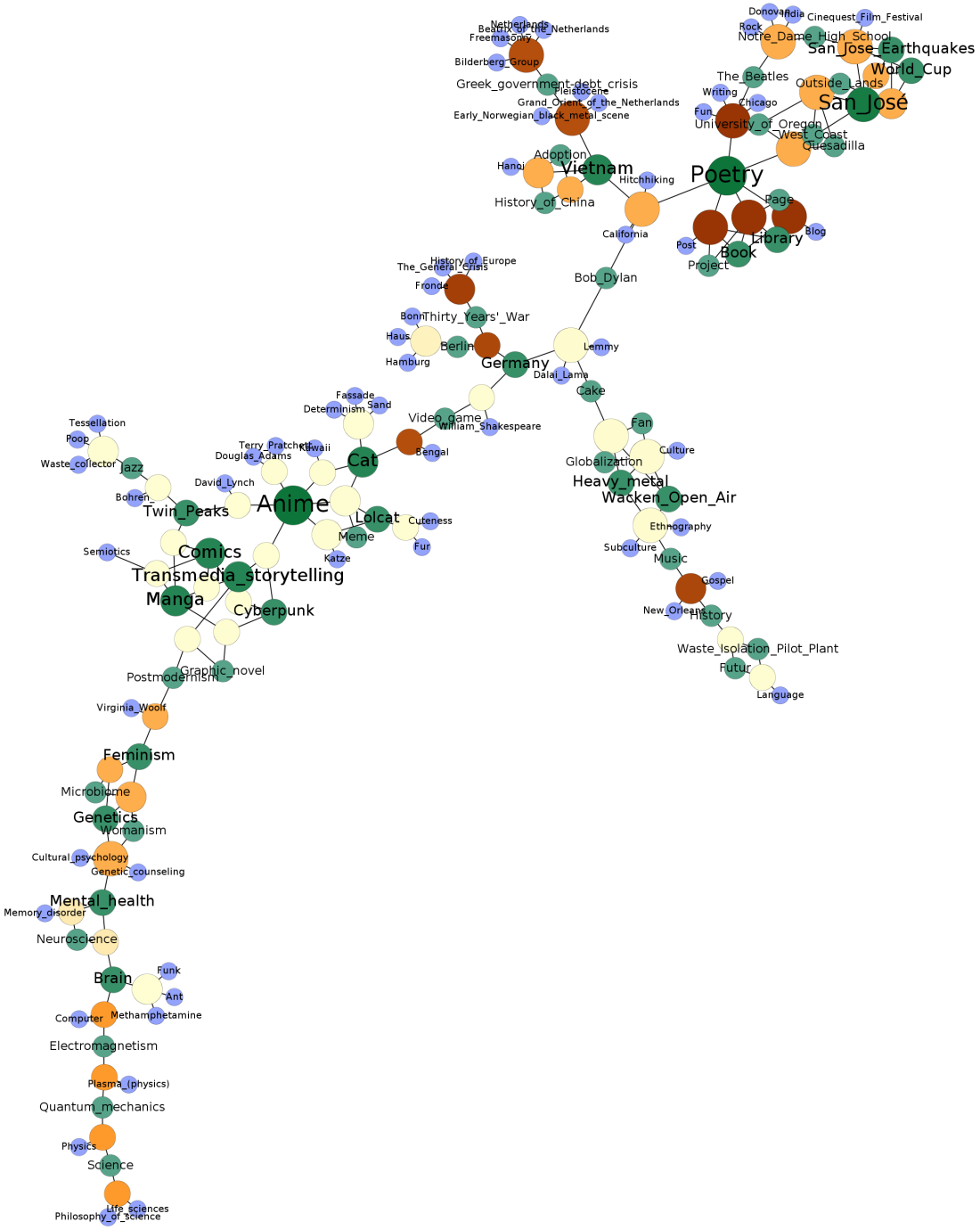


Abbildung 5.13: *Netz von Visualisierungen verbunden über die gewählten Ausgangsartikel des Betrachtungszeitraums, die größte Komponente.*

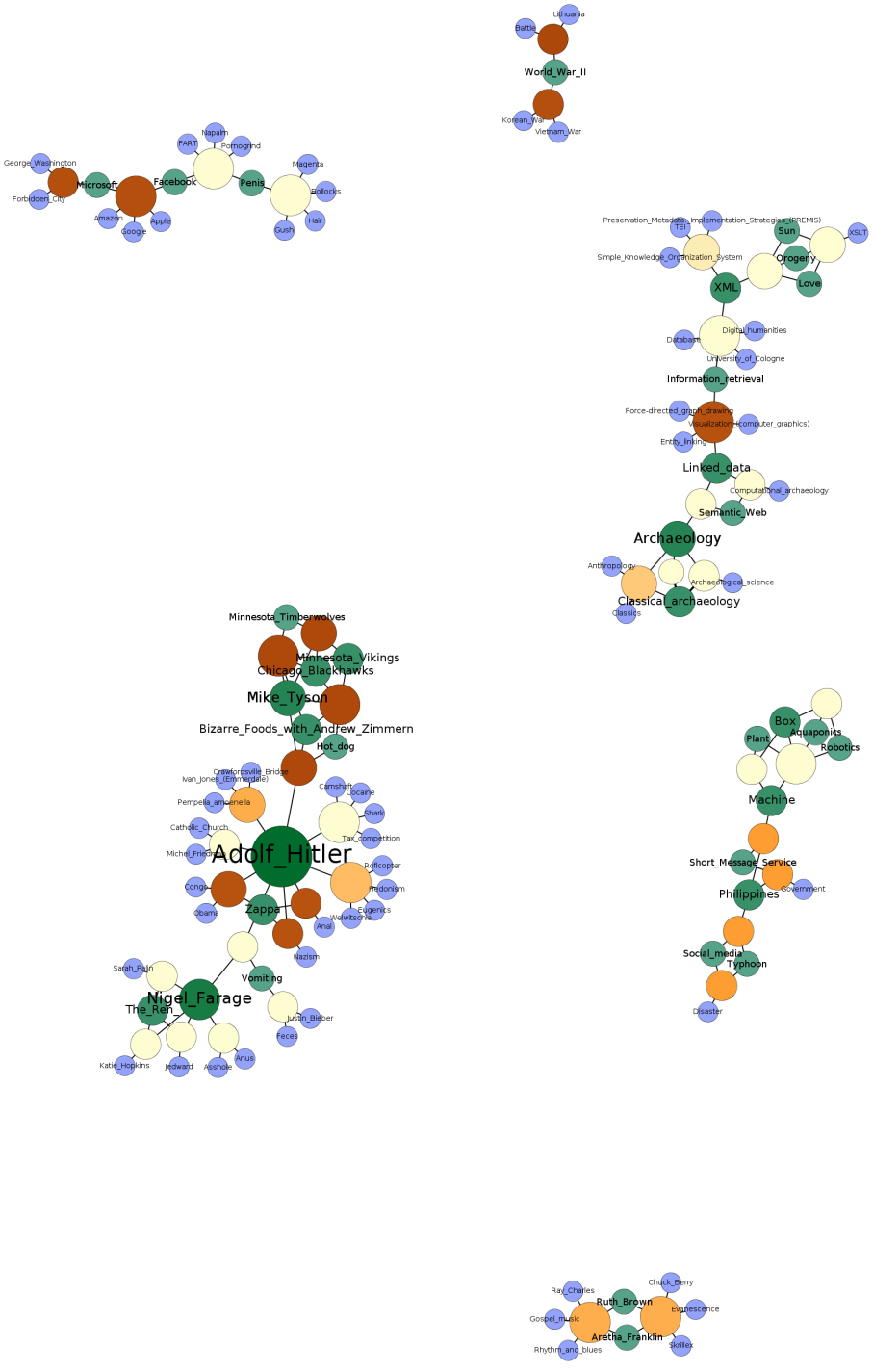


Abbildung 5.14: *Netz von Visualisierungen verbunden über die gewählten Ausgangsartikel des Betrachtungszeitraums, eine Auswahl kleinerer Komponenten.*

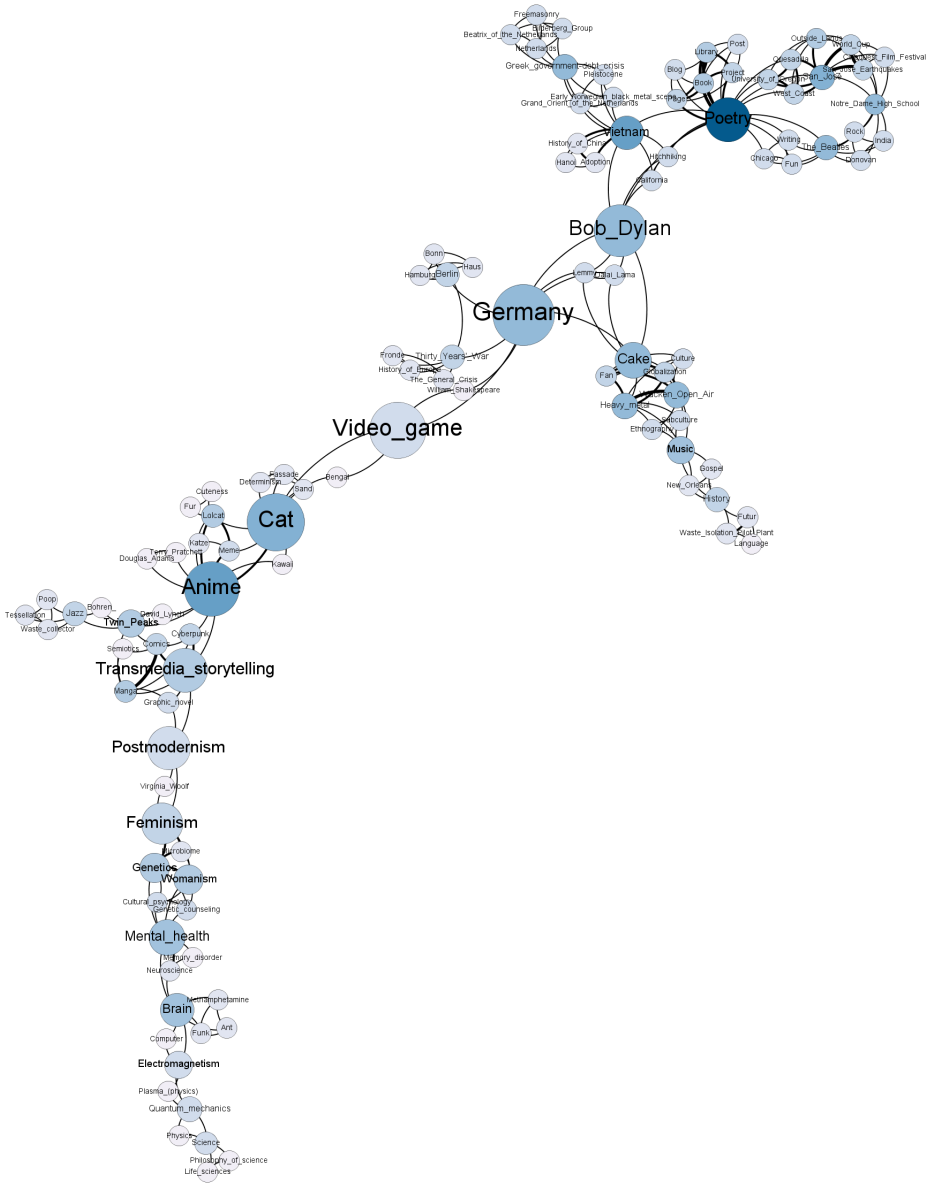


Abbildung 5.15: *Projektion der größten Komponente aus Abbildung 5.13. Die Farbintensität entspricht der lokalen Wertung (Grad) die Größe der Knoten entspricht der transitiven Wertung(Betweenness Centrality).*

Analysen des Nutzungsverhaltens

Im vorherigen Abschnitt wurden die erstellten Visualisierungen und die Ausgangsartikel untersucht. Eine ähnliche Analyse kann auch auf das genauere Nutzerverhalten in den Visualisierungen selbst vorgenommen werden.
Die Klicks der Nutzer auf diese Visualisierungen können in einer zusammenhängenden Betrachtungsweise visualisiert werden. Hierzu werden die angeklickten Knoten mit den Visualisierungen, in denen sie angeklickt wurden, abgebildet. Dabei sticht heraus, was visualisierungsübergreifend interessant war und wie sich die Interessensgebiete zusammensetzen können.
Es werden zwei Arten von Knoten verwendet:
Erstens die Artikel, die in Visualisierungen angeklickt wurden. Sie sind in der folgenden Abbildung blau gefärbt.
Zweitens die Visualisierungen, in denen die Knoten angeklickt wurden. Diese sind in den folgenden Abbildungen rot und ohne Beschriftung. Handelt es sich um ein Beispiel, dann sind die Knoten grün und beschriftet.
Diese beiden Knotentypen bilden für sich kleine sternförmige Komponenten. Diese werden durch die Knoten zusammengehalten, die in verschiedenen Visualisierungen angeklickt wurden.
Das sich hieraus ergebende Bild zeigt exemplarisch, wie sich welche Cluster bilden und welche Knoten visualisierungsübergreifend angeklickt wurden.
Im Folgenden werden die drei größten Komponenten dieser Zusammenhänge beschrieben. Dabei werden nur Knoten oder Artikel angezeigt, die in mindestens zwei verschiedenen Visualisierungen angeklickt wurden. Die Größe der Knoten und Labels wird nach der Betweeness Centrality vergeben. Diese Gewichtung hebt transitiv wichtige Knoten hervor.
In der größten Komponente, Abbildung 5.16, spiegeln sich Themenbereiche wieder. Hier befinden sich auch viele Beispiele, um die sich mehrere andere Knoten angesiedelt haben.
In der zweitgrößten Komponente, Abbildung 5.17, werden zwei Komponenten durch den Artikel “Israel” verbunden. Dieser kommt im oberen Themenbereich zum Rechtsterrorismus in Deutschland und im unteren Bereich im Zusammenhang mit Musik und Computern vor.
Die drittgrößte Komponente, Abbildung 5.18, bildet sich um die Beispiele der Science-Fiction Filme und den Beispielen zum Posthumanismus. Dabei fällt der Autor Phillip K. Dick ins Auge. Er nimmt eine zentrale Rolle zwischen den beiden Clustern ein, da er durch die Popularität seiner Werke offenbar ein recht guter Orientierungspunkt ist. Science-Fiction selbst als zentralster Knoten, beschreibt die Überschneidungspunkte.
In diesen Visualisierungen werden Elemente und Bereiche von Interesse aufgedeckt. Die angebotenen Beispiele lassen sich hier direkt erkennen. Doch ist der Zusammenhang von Komponenten sehr stark von der Interaktion des Nutzers abhängig.
In der größten Komponente ist einer der zentralsten Knoten mit den Vereinigten Staaten (oben rechts) besetzt. Dies ist der am häufigsten verlinkte Artikel in der Englischen Wikipedia. Daher ist die Wahrscheinlichkeit groß, dass der Artikel oft auftaucht und oft geklickt werden kann. Jedoch wird dieser Artikel immer sehr klein abgebildet werden, da

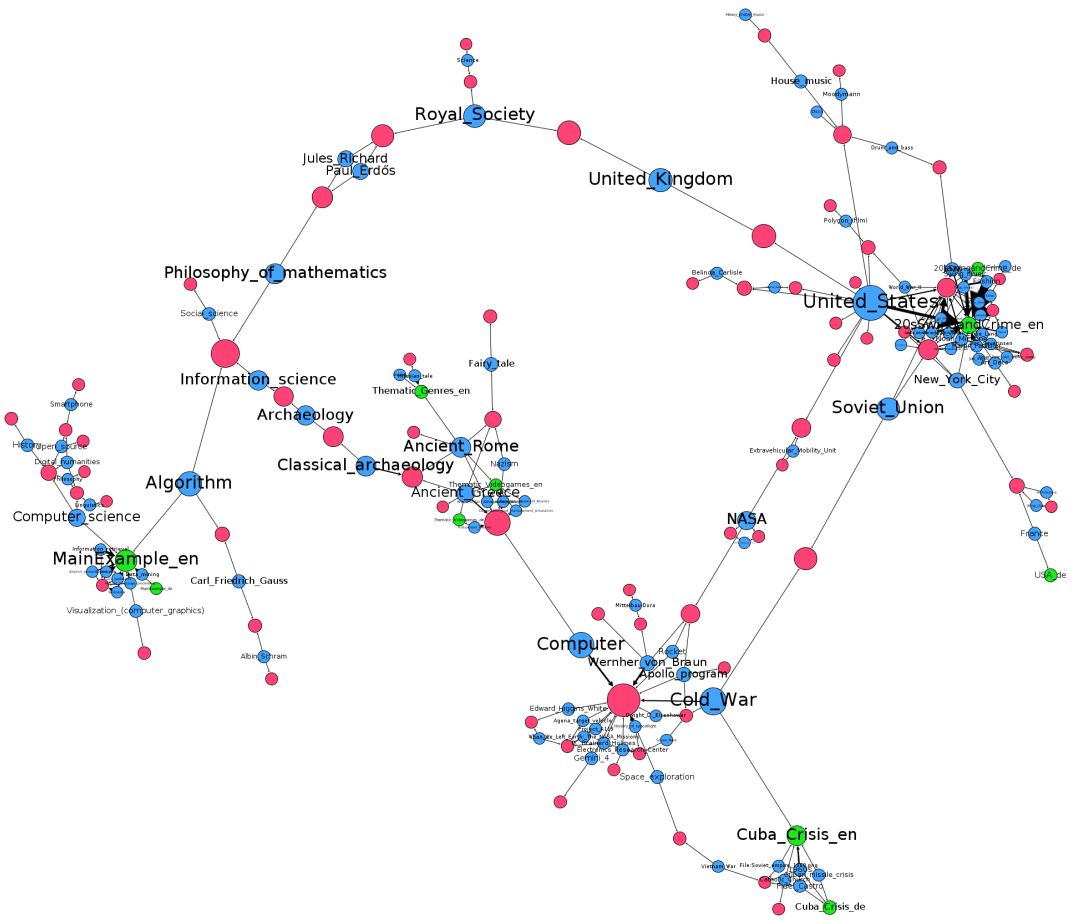


Abbildung 5.16: *Zusammenhänge über Klicks, größte Komponente*

keine der hier erstellten Visualisierungen das Umfeld des Artikels auch nur annähernd abdecken kann. Auch sind viele der Knoten an zentralen Brückenpositionen Länder, welche im generellen stark in Wikipedia verlinkt sind.
Die in diesem Abschnitt gezeigten Visualisierungen zeigen das Klickverhalten der Nutzer und stellen die Gesamtheit der Daten in größerem Zusammenhang dar.

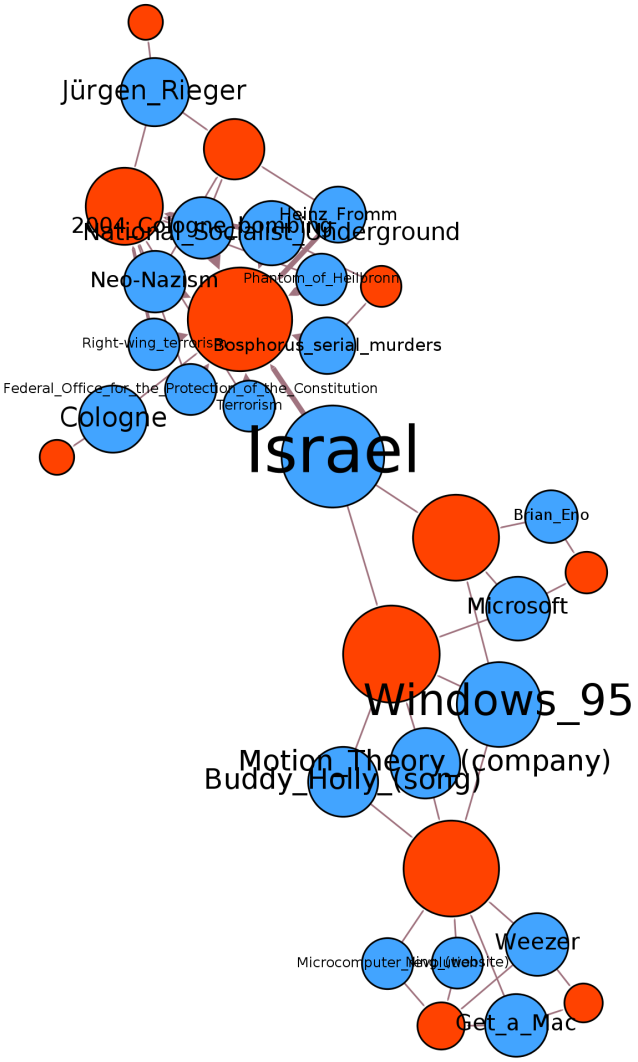


Abbildung 5.17: *Zusammenhänge über Klicks, zweitgrößte Komponente*

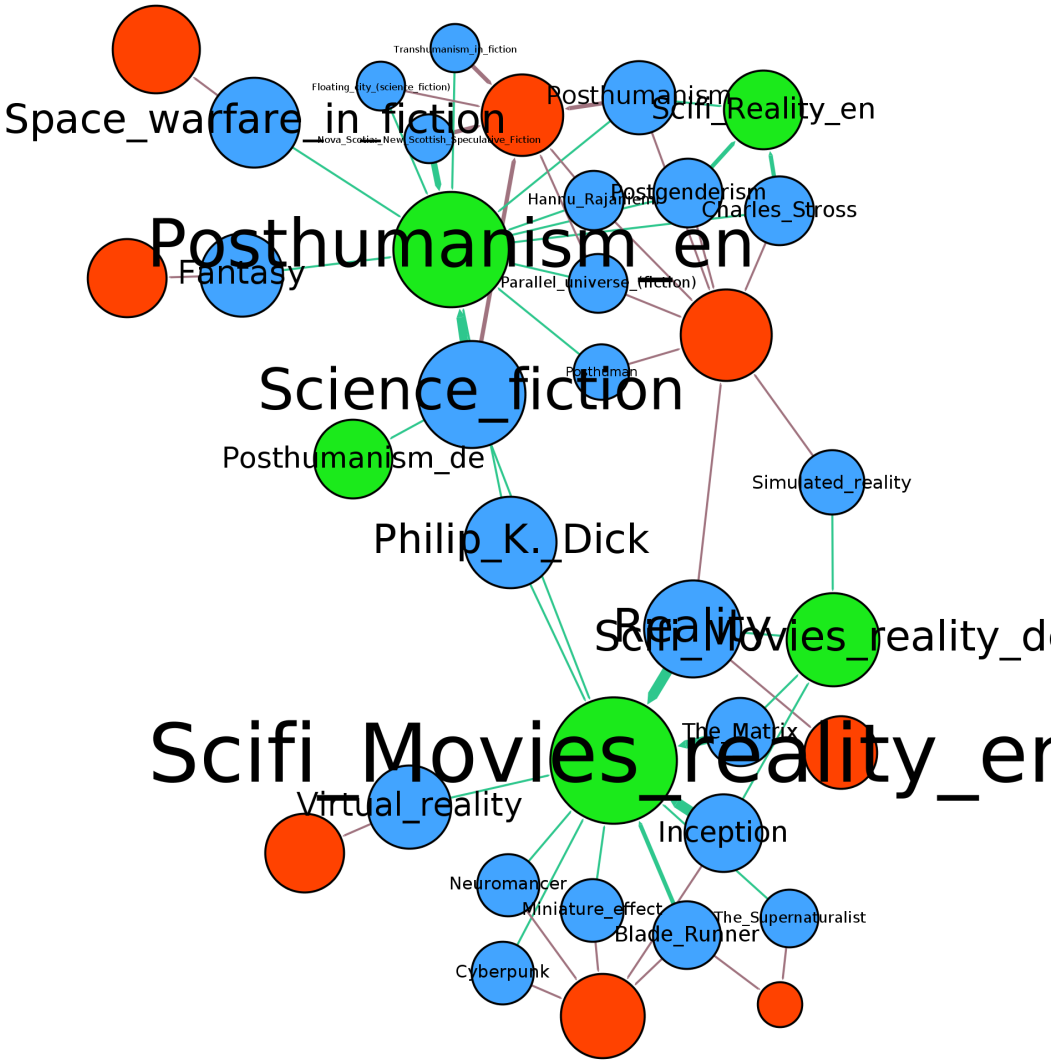


Abbildung 5.18: *Zusammenhänge über Klicks, drittgrößte Komponente*

5.4.4 Auswahl und Vergleich

Da eine entitätsbasierte Auswahl getroffen wurde, um eine Visualisierung zu erstellen, ist es interessant zu wissen, welche Art von Knoten von den Nutzern dafür verwendet wurden. Dafür werden drei verschiedene Arten von Knoten betrachtet, die Ausgangsartikel, die einfach geklickten Artikel und die geöffneten, also doppelt geklickten Artikel.
Die Visualisierungen wurden aufgrund der Eingangsmenge von 3-5 Artikeln erstellt.
Der Grund für diese Auswahlen liegt im explorativen Vorgehen der Nutzer. Es geht um die benennung von Bekanntem um Unbekanntes zu finden. Eine Auswahl Ad-Hoc zu benennen, ist für den Nutzer in diesem Retrievalansatz die größte Herausforderung.
Es wird im Folgenden angenommen, dass bekannte und direkt beschreibbare Entitäten als Ausgangsartikel verwendet wurden. Dem Nutzer bekannte Dinge können auch durch die Konstruktion der Wikipedia beschrieben werden. Um etwas benennen zu können, muss es bekannt sein. Das Gleiche gilt, wenn etwas in Wikipedia verlinkt wird.
Es wird Angenommen, dass Entitäten, die benannt werden können, einen hohen eingehenden Grad in der Wikipedia haben.
Bei der Interaktion mit Knoten in der Visualisierung gibt es zwei verschiedene Möglichkeiten, die Auswahl mit einfachem Klick: Das Fokussieren auf den Bereich um einen Artikel.
mit dem zweiten Klick: Das Öffnen des beschriebenen Artikels in einem neuen Browser-Tab.
Die Auswahl mit einfachem Klick ist dabei eine Orientierungshilfe. Sie hebt das Umfeld eines Datensatzes hervor, das Ego-Netz um den ausgewählten Knoten. Daher wird angenommen, dass ein einfacher Klick ohne zweiten Klick eine gewisse Vertrautheit mit dem Artikel angibt, da er zum weiteren Filtern verwendet wird.
Die Auswahl mit Zweitem Klick zeigt etwas Neues an. Da unter einem Begriff nicht direkt verstanden werden kann, was gemeint ist, wird zur Klärung hier der Artikel verwendet.
Daraus ergibt sich die Annahme:
Ein Knoten, der einfach geklickt wurde, hat eine höhere Bekanntheit als ein Knoten, der zweimal geklickt wurde.
Daraus folgt die Annahme, dass ein einfach geklicker Knoten einen höheren In-Degree besitzt als ein doppelt geklickter Knoten, der bekannt ist.

Auftrittshäufigkeit

Die drei beschrieben Kategorien von Artikeln werden anhand ihres eingehenden Grades miteinander verglichen.
Da es sich um Mengen verschiedener Größe handelt, muss eine Vergleichsmöglichkeit für die Verteilung des eingehenden Grads gefunden werden. Mit Perzentilen lässt sich die Verteilung des eingehenden Grades dieser verschieden großen Mengen vergleichen. Ein Perzentil beschreibt, wie hoch der maximal Wert für einen Prozentsatz ist. Bei einem Perzentil von 30 Prozent und einem Wert von 35 heißt das: 30 Prozent der Knoten haben einen eingehenden Grad von unter 35.
Mit diesem Vorgehen ist die Analyse unabhängig von der Menge der untersuchten Knoten. Wie in Abbildung 5.19 zu sehen ist, haben die bei der Erstellung der Visualisierung

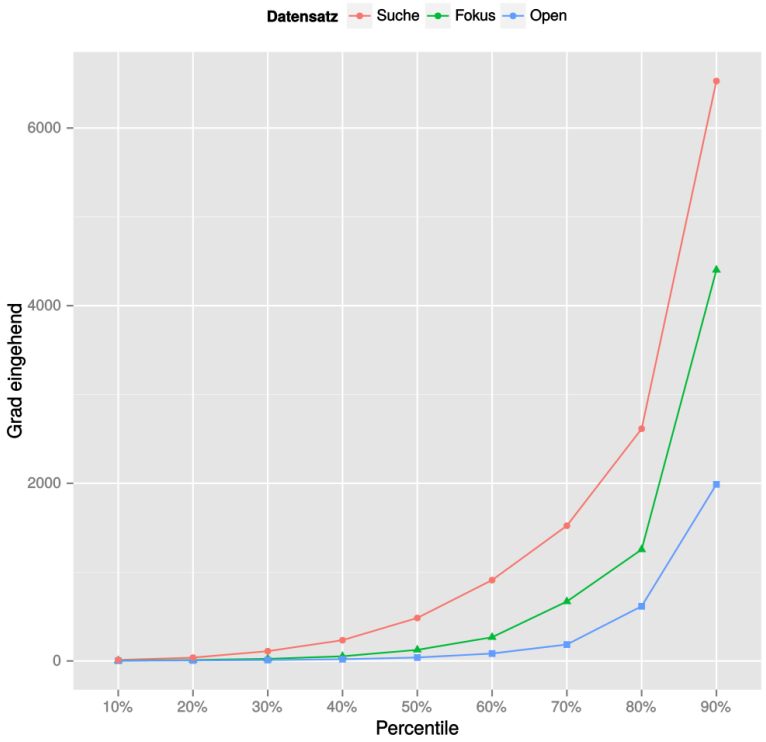


Abbildung 5.19: *Vergleich der eingehend Gradzahl in den Ausgangsknoten und den einmalig (Fokus auf das direkte Umfeld) sowie zweimalig (Öffnen des Artikels) geklickten Knoten.*

ausgewählten Ausgangsartikel die höchsten Werte beim eingehenden Grad. Das heißt, die Nutzer konnten aus dem Stehgreif nur bekannte Wiki-Artikel benennen.
Die fokussierten Knoten haben wie angenommen einen höheren eingehenden Grad als die geöffneten Artikel. Sie scheinen daher bekannter zu sein. Bei den ausgehenden Knoten lassen sich ähnliche Verhältnisse beobachten, nur ist hier der Unterschied nicht so groß, da die Menge der ausgehenden Links von der Größe und damit der Ausführlichkeit eines Artikels abhängt.

Tabelle 5.1: *Statistische Unterschiede zwischen Auswahl- und Zufallsauswahl*

Typ	Auswahl	Zufall	Auswahl/Zufall	% Auswahl/Zufall
Nscources	3.5965	3.3054	0.2910	8.80%
Anzahl_Knoten	86.4523	29.7059	56.7464	191.03%
Anzahl_Kanten	669.4568	114.9395	554.5172	482.44%
Hidden_Context	48.7561	14.8168	33.9393	229.06%
Hidden_Context_/_Kontextknoten	0.4605	0.4069	0.0536	13.17%
CWS	8.0909	6.7450	1.3459	19.95%
CWS_pro_Kontextknoten	0.2466	0.5139	-0.2673	-52.01%
Filter	0.8359	0.0825	0.7534	913.31%

Mensch gegen Zufall

In den vorhergehenden Visualisierungen war auffällig, dass bei den Ausgangsartikeln eher allgemeine Artikel verwendet wurden. Sie scheinen recht bekannt zu sein. Gerade die zentralen Positionen wurden von Ländern eingenommen. Bei den angeklickten Artikeln waren auch viele bekannte Artikel an zentralen Positionen. Die Artikelbezeichner waren jedoch augenscheinlich länger und haben sich gelegentlich überschnitten. Die geklickten Artikel scheinen spezieller zu sein, als jene, die als Augangsknoten gewählt wurden.
Es ist davon auszugehen, dass Nutzer Artikelkombinationen wählten, die intentional zusammengestellt wurden. Im Weiteren wird untersucht, ob die Ergebnisse aus dem LWMap Service von den Eingangsknoten abhängen und ob eine intentionale Auswahl Unterschiede zu einer zufälligen Auswahl aufweisen. In statistischer Hinsicht werden Netze untersucht, indem zufällige Netze mit möglichst ähnlichen Eigenschaften erzeugt werden wie die zu untersuchenden Netze. Gängig ist es zufällige Netze mit der gleichen Anzahl von Knoten und Kanten zu erzeugen. Als erweiterndes Kriterium kann beispielsweise auch die Gradverteilung aus dem zu untersuchenden Datensatz nachgebildet werden.
Zur Abgrenzung vom Zufall im LWMap Service werden zufällige Netze erzeugt, bei denen die Anzahl der Ausgangknoten der Verteilung im gemessenen Zustand entspricht. Des Weiteren werden nur Knoten verwendet, die auch von den Nutzern als Ausgangsknoten verwendet wurden. Weiterhin wird verhindert, dass sich die zufälligen Zusammenstellungen der Knoten den Originalzusammenstellungen gleichen oder sich stark mit diesen überschneiden.
Nach diesen Grundlagen wurden vier Zufallsstichproben gezogen, die dem untersuchten Datensatz entsprechen.
Generell sind die Unterschiede zwischen den Zufallsstichproben und der Auswahlstichprobe sehr groß, wie in Tabelle 5.1 zu sehen ist. Die ursprüngliche Menge an erfolgreichen Visualisierungen, die ein Ergebnis erzeugten, liegt bei 451. Die vier zufälligen Stichproben haben 391, 390, 406 und 402 erfolgreiche Visualisierungen. Die Zufallszusammenstellungen haben also um etwa 10 % weniger erfolgreiche Darstellungen. Am Auffälligsten sind Werte, die sich auf die Anzahl der Knoten und Kanten beziehen. Dabei ist die Anzahl der Kanten um fast das Sechsfache (482,44 % mehr) höher und die Anzahl der Knoten um ungefähr das Dreifache (191,03 % mehr) höher als in den Zufallsstichproben. Die Anzahl der Kanten unterscheidet sich bei der Menge der Kontextkanten am stärksten. Hier gibt es durchschnittlich die Siebeneinhalbfache Menge (656,41 % mehr) an Kanten in der Auswahlstichprobe, als bei den Zufallsstichproben.
Da die Netze sich in ihrer Ausprägung sehr stark unterscheiden, ist eine Normalisierung nötig. In der Betrachtung des Hidden Context, also der Knoten die nur mit ausgehend gerichteten Kanten mit den Ausgangsknoten verbunden sind, wird eine Scheinauffälligkeit sichtbar. Insgesamt hat die Auswahlstichprobe 229,06 % mehr Hidden Context-Knoten als die Zufallsstichprobe. Normalisiert über die Menge der Kontextknoten in einem Graphen nähern sich die Werte jedoch an und die Auswahlstichprobe hat nur 13,17 % mehr Hidden Context pro Kontextknoten. Dieser Wert scheint, relativ gesehen, keinen allzu großen Unterschied zu machen.

Tabelle 5.2: *Auftreten der symmetrischen Verbindung zwischen Knoten des Typs Place,PopulatedPlace,Country*

Menge	Ranglistenplatz	Absolut	Prozentual
Auswahl	31	955	0.45823577
Zufall1	2	856	3.42962458
Zufall2	3	553	2.43290805
Zufall3	4	645	2.45125983
Zufall4	3	826	3.58958759

Interessant ist auch, inwieweit ein gemeinsamer Zusammenhang der Kontextknoten existiert. Ein Indikator dafür ist die Anzahl der Komponenten, in die der Gesamtgraph zerfällt, wenn die Ausgangsknoten entfernt werden. Absolut gesehen zerfallen Auswahlgraphen um 19,95 % stärker. Bereinigt um die Anzahl der Kontextknoten zeigt sich jedoch, dass die zufälligen Graphen prozentual stärker zerfallen als die Auswahlgraphen. Die Auswahlgraphen zerfallen dabei durchschnittlich nur in die Hälfte der Komponenten als die zufälligen Graphen.
Um die Abweichungen genauer zu klären wurden die Verbindungen zwischen den Typen von Artikeln gemessen. Dabei waren beidseitig untypisierte Verbindungen in allen Proben am häufigsten. Darauf folgen die einseitig typisierten Kanten in der Rangliste. Es ist bei allen zufälligen Ziehungen auffällig, dass die Stichproben große Mengen einer Endpunktkombination aufwiesen, die typisierten symmetrischen Verbindungen:
Place, PopulatedPlace, Country - Place, PopulatedPlace, Country
Das gilt für die Rangliste der häufigsten Kombinationen, sowie für den relativen Anteil an allen Kanten, wie in Tabelle 5.2 gezeigt wird. Absolut weist die Nutzer-Stichprobe sehr viel höhere Werte auf, aber da sie generell sehr viel mehr Knoten als die Zufallsstichproben aufweist, ist der absolute Wert irreführend.
Die Ranglistenplätze und der prozentuale Anteil an der Gesamtmenge der Knoten zeigt, dass dieser Verbindungstyp bei den Zufallsstichproben öfter vertreten ist, als bei der Auswahlmenge.
Die Nutzer scheinen ein bestimmtes Wissensfeld im Kopf zu haben. Dies wird durch die Angabe von Ausgangsartikel umrissen und lässt sich in den Daten wiederfinden.
Ein überproportionales Auftreten von Kanten zwischen Ländern scheint dabei ein recht zuverlässiger Indikator für eine Zufallsauswahl zu sein und erfolgreiche Visualisierungen bei den Zufallsstrichproben zu erzeugen.

5.4.5 Resümee

Einen Onlineservice zu entwickeln, zu betreiben und auszuwerten ist eine recht umfangreiche Aufgabe. Gerade die selbstgewählten theoretischen Auflagen zur Nachhaltigkeit haben zu Kompromissen geführt.
Dazu gehört beispielsweise die Referenzierbarkeit der Wikipedia per link und Referenzierbarkeit der Filter und Ansichten im LWMap Service. Die Knoten des LWMap Service referenzieren immer auf die aktuellste Version des angeführten Artikels. Die sauberere Form wäre, die präzise OldID anzugeben, die die Version des Artikels referenziert zu der

die DBPedia Daten erstellt wurden. Die genaue ID ist dabei nur schwer zu ermitteln da die DBPedia keinen sehr genauen Zeitpunkt für die Datenherkunft angibt.
Die Referenzierbarkeit auf eine Ansicht ist problematisch, da ein sehr langer für Menschen unlesbarer Link entstand.
Es war wichtig einen Kompromiss in der Nachhaltigkeit zu finden. Dabei konnte nicht auf die Versprechen von nachhaltiger Infrastruktur gewartet werden.
Selbst wenn eine solche nachhaltige Infrastuktur Gewissheit wäre, würde die Auswahl der technologischen Prämissen der Betreiber immer die Freiheit derer einschänken, die diese Infrastruktur nutzen wollen.
Da die Bereitstellung der Ergebnisse bis nach der Veröffentlichung aus verfahrenstechnischen Gründen warten muss und Nachhaltigkeit ein Langzeit Versprechen ist, kann der LWMap Service strenggenommen nur ein unvollendetes Experiment bleiben.
Ein kritischer Punkt, der bei solchen Experimenten angeführt werden muss, ist die Aufmerksamkeit, wie in der Einleitung dieser Arbeit beschrieben. Im Internet eröffnet sich dem potenziellen Nutzern eine ganze Welt von Möglichkeiten und mit wissenschaftliche Projekten genug aufmerksamkeit zu erzeugen, ist schwierig.
Hier ist auch das Argument für eine qualtitive Datenauswertung und eine deskriptive Statistik zu finden. Eine kritische Masse an Zugriffen für eine valide inferenzbasierte Statistik zu sammeln, ist schwer zu bewerkstelligen.
Dennoch musste in der Auswertung vom einzelnen Nutzer abstrahiert werden. Dies wurde u.a. mit der Umwandlung der Zeitdaten in eine Abfolge von Darstellungen auf den Datensätzen versucht.
Die Auswertung der Nutzungsdaten hat dabei den Bias der Nutzer zutage gefördert. Die hier zusammengefundenen Komponenten scheinen nicht intentional entstanden zu sein. Dabei haben sich Nutzer auch an den Beispielen orientiert und diese abgewandelt.
Dass die zusammengestellten Auswahlen trotzdem funktionieren, zeigen die Ergebnisse im Vergleich mit den zufälligen Zusammenstellungen. Es könnte angenommen werden, dass es bei der Dichte der Wikipedia-Page-Links egal ist, welche Knoten als Eingangsknoten funktionieren. Dies wurde widerlegt. Auch wenn die verwendeten Artikel einen höheren eingehenden Grad besitzen, sind die Ergebnisse bei der zufälligen Zusammenstellung sehr viel kleiner.

5.5 Diskussion

In dieser Arbeit wurde gezeigt, wie mit netzwerkanalytischen Verfahren im Umfeld der Geisteswissenschaften gearbeitet werden kann. Dabei wurden vor allem text-, kategorie- und entitätslastige Datenstrukturen untersucht.
Bei komplexen, multipartiten Datensätzen ist das Auffinden von interessanten Perspektiven und Ausschnitten wichtig um einen Einblick in die Daten und ihre Struktur zu erhalten. Bei Netzen, die verschiedene Typen von Entitäten verbinden, können die entstehenden Muster komplexe Sachverhalte abbilden. Nicht immer machen die gefundenen Netzwerke direkt Sinn oder aber es überlagern sich verschiedene Informationsstrukturen. Eine Ansicht, die über die einzelnen Datenfragmente hinausgeht, zeigt dem Experten den

Umfang, die Vollständigkeit und den Fokus eines Datensatzes oder eines Datenausschnitts. Dem Laien wird der Einstieg vereinfacht und eine zermürbende Daten-Lektüre erspart.
Das generelle Netzmodell wurde in dieser Arbeit als Startpunkt für einzelne Untersuchungen gesetzt. Dabei wird über den Betrachtungszeitraum das offene Modell durch die konkret betrachteten Daten strukturell eingeschränkt. Das erarbeitete Modell sollte sich nicht nur aus der Netzstruktur selbst, sondern aus anderen Eigenschaften der Daten formen.
Wie in der Einleitung beschrieben, gibt es immer mehr Daten und diese Daten ziehen Daten nach sich. In Abschnitt 5.4.3 wurde gezeigt, wie die Nutzung von Daten neue Daten erzeugt.
Mit dem Aufkommen von maschinellem Lernen und dem damit verbundenen Lernen aus Daten wird die Datenqualität immer wichtiger. Trainingsets und deren logischer zusammenhang wird ein Aspekt sein der Manueller Prüfung bedürfen wird. Ein anderer Aspekt sind die Entscheidungen und Handlungen der trainierten künstlichen Intelligenz, die Ausgabedaten. Diese diese müssen Untersucht werden um die Handlungen und Zusammenhänge die Erzeugt werden untersuchen zu können.
Natürlich sind Netze allein nicht die Antwort auf die Datentechnischen Herausforderung. Sie sind vielmehr ein Teilaspekt, wie mit Daten umgegangen werden kann und sollten mit anderen Techniken kombiniert werden. Der Vorteil des graphenbasierten Vorgehens liegt an der qualitativen, übergreifenden Untersuchung von Datensätzen über die einzelnen Datenfragmente oder deren quantitativer Gewichtung hinaus.

5.5.1 Visualisierungen

Visualisierungen von Netzen sind eine Möglichkeit, Zugang zu einem komplexen Datensatz zu schaffen, im speziellen interaktive Visualisierungen sind ein wichtiger Teilaspekt dieser Arbeit. Im Folgenden werden die wichtigsten Erfahrungen aus dieser Arbeit zusammengetragen.

Anordnung von Knoten

Eine generelle, vollautomatische Lösung für die Knotenanordnung mit einem einzigen Algorithmus führt zu nicht zufriedenstellenden Lösungen. Das Wissen über über den Aufbau der zu visualisierenden Daten ist wichtig, um ein möglichst gutes Ergebnis zu erzielen.
Kräftebasierende Verfahren sind nicht in der Lage beliebige Mengen von Knoten und Kanten lesbar zu visualisieren. Hier können einzelne unparametrisierte Algorithmen keine gute Visualisierung schaffen. Beispielsweise erzeugt zuviel Anziehung zwischen Knoten unübersichtliches Verklumpen von Netzen, es entstehen die sogenannten Hairballs.
Um Hairballs zu Vermeiden müssen die richtigen Parameter und Grundordnungen auf die Daten angewendet werden. Es können auch Filter verwendet werden, um Aspekte sichtbar zu machen.
Wie ein solches Vorgehen aussehen kann, wurde anhand des LWMap Service im Detail beschrieben (Abschnitt 5.3.5). Die Parameter wurden unter anderem an die Menge der

Knoten und Kanten angepasst. Die ganze Anordnung der Knoten ist jedoch ein komplexer Vorgang, der sich nah an der Struktur der Daten orientiert.
Netzwerk Visualisierungen sollten möglichst großflächig dargestellt werden. Im gegensatz zu Balkendiagrammen ist die Informationsdichte meistens sehrviel höher. Diese Strategie ist natürlich nur erfolgreich, wenn der Platz auch genutzt wird und nicht ein Großteil der Knoten wenig Platz in der Mitte der Darstellung einnimmt.
Um mehr Daten zugänglich zu machen, kann Interaktivität wie das Zoomen und dynamisches Filtern eingesetzt werden. Dabei können Knoten oder deren Beschriftungen erst auf hohen Zoomstufen eingeblendet werden. Filter können auch die Dichte eines Graphen reduzieren, sodass interessante Strukturen besser sichtbar werden.
Dem Paradigma des interaktiven Anordnens der Knoten durch den rezipierenden Nutzer, wie es öfters in interaktiven Graphenvisualisierungen zu finden ist, wurde in dieser Arbeit eine klare Absage erteilt. Dies geschah aus mehreren Gründen:
1. Um eine diskursive Visualisierung zu erhalten. Alle Rezipienten müssen das Gleiche sehen, um eine diskursive Grundlage zu haben.
2. Das Problem, eine sinnvolle Ordnung zu finden, wird in Richtung des Nutzers verschoben. Der Bereitsteller kennt die Daten und sollte dieses Problem daher besser lösen können als der Nutzer.
3. Durch die willkürliche Anordnung wird der letzte Rest algorithmischer "Neutralität" entfernt. Der Nutzer kann seine eigenen Vorstellungen in den Daten einbetten und suggestiv seine eigene Meinung bestätigen.
Mit dem LWMap Service wurde eine Fragestellungen an die Daten erarbeitet. Das Ergebnis wird jedem Nutzer in der gleichen Weise angezeigt. Dem Nutzer wird dabei jedoch durch die Bereitstellung der Quelldaten die Möglichkeit gegeben, alternative Visualisierung zu erstellen. Die Neutralität bezieht sich hier darauf, dass eine ganze Klasse von Graphen gleich visualisiert wird.

Grundordnung

Netzwerkvisualisierungen haben im Gegensatz zu geografischen Anordnungen das Problem, dass sie keinen klaren definierten Raum abbilden. In mehreren ähnlichen Abbildungen gibt es daher wenig Orientierungsmöglichkeiten für den Rezipienten.
In dieser Arbeit wurde daher versucht, die Anordnung der Knoten über Darstellungen hinweg immer möglichst konstant zu halten. Dadurch wurde den Rezipienten eine möglichst einfache Orientierung erlaubt. Wenn Knoten gefiltert (Abbildungen in Abschnitt 3.2.1) oder Kanten neu klassifiziert wurden (Abbildungen: 4.3, 4.4,4.5) blieben die Strukturen in allen Darstellungen an den gleichen Plätzen. Entstehende Neuvisualisierungen müssen nicht zwangsweise die absolut gleiche Anordnung aufweisen. Die Visualisierungen können auch aufeinander aufbauen, wie in Abschnitt 3.2.2 an den Visualisierungen gezeigt wurde. Die zirkuläre Abbildung 3.8 diente als Grundlage für die Anordnung in 3.9 und diese für das Detail in Abbildung 3.10. In Abschnitt 3.3 wurden die Abbildungen 3.12b, 3.13b und 3.14 so gedreht, dass die Wahlpräferenzen immer in der gleichen Orientierungsrichtung

lagen.

Farbe und Metainformationen

Auch die Art der Knoten, die betrachtet wird, ist kritisch, da zu viele Typen von Knoten eine Visualisierung komplex machen. Das spricht gegen den Zweck von Visualisierung zur Komplexitätsreduktion und Anpassung an das menschliche Sinnsystem.
Zur Färbung können auch Attribute genutzt werden die normalerweise die Position bestimmen, wie relative Zeitpunkte oder Reihenfolgen. Zu nennen sind hier die Akte bei Macbeth (Alle Abbildungen in Abschnitt 3.2.1) der Abfolge von Paragrafen (Abbildung: 3.9) oder der Erstellungsabfolge im LWMap Service (Abbildungen: 5.13 und 5.14).
Auch ist die Wahl einer möglichst kurzen und knappen Beschriftung von Knoten wichtig, denn auch diese Beschriftungen können überlappen und unlesbar werden. Im Fall der DBpedia war die Beschreibung der Entitäten meist kurz und prägnant. Das ist der guten Kuratierung der Wikipedianer zu verdanken. In anderen Datensätzen führt das zu Problemen wie bei den Visualisierungen zur Arachne (Abschnitt 3.6) oder den Gesetzestexten (Abschnitt 3.2.2). Die Frage der individuellen Identifizierbarkeit ist aber auch von der Fragestellung abhängig. Muss jeder Knoten direkt einzeln identifiziert werden oder reicht die Abbildung einer Kategorie durch eine Farbe? Wenn die Darstellung interaktiv ist, dann kann auch eine Dereferenzierung per Link auf einen Datensatz sinnvoller sein.

Kochrezepte

In dieser Arbeit wurde vorgestellt wie ein parametrisierbares Script oder auch “Kochrezept” zur Netzwerkvisualisierungen entwickelt werden kann. Dadurch soll eine Klasse von Daten unter berücksichtigung der Knoten- und Kantenmenge automatisch als Netzwerk visualisiert werden können. Diese Klasse von Daten umfasst einzeldaten aber auch Ergebnis von Abfragen auf einem größeren Datenbestand. Um die in dieser Arbeit verwendeten Lösungen zu entwickeln, war manuelle Parametrisierung der Algorithmen nötig. Hierfür wurde größtenteils die Software Gephi [18] verwendet, da die im Frontend entwickelten Vorgänge mithilfe der Kernbibliothek in ein Script überführt werden können.
Zum konzept “Kochrezept” gehören aber auch die Verbindung einzelner Schritte wie Vorordnung, Fixierung, Aufteilen und die sukzessive Anordnung der Daten. Diese Schritte sind bei Macbeth (Abschnitt 3.2.1), den DBpedia Rivalitäten und Allianzen (Abschnitt: 4.3.2 Abbildungen: 4.3, 4.4,4.5) sowie dem LWMap Service (5.3) zum Einsatz gekommen. Die Anpassung der Kantengewichte wie bei LWMap Service (Abschnitt: 5.3.4) ist eine ergänzende Strategie.
Das Zusammenwirken und Abbilden dieser Schritte in einem Script wäre demnach das beschriebene “Kochrezept”. Da auch Kochrezepte nicht immer in jedem Fall die besten Ergebnisse liefern, sollten die Daten immer so frei sein, dass ein Gegenentwurf angefertigt werden kann.

Nachhaltige Darstellung

Wie in dieser Arbeit gezeigt wurde, ist die nachhaltige und nachvollziehbare Bereitstellung von Visualisierung problematisch. Viele Datenbestände sind Work in Progress. Das erschwert eine nachhaltige Referenzierbarkeit der Quellen und deren Zustand.
Auch lassen sich die Änderungen nicht immer zu jedem Zeitpunkt reproduzieren. Das Zeitproblem zeigt sich auch bei der Interpretationstechnologie. Eine interaktive Darstellung muss auf einem interaktiven System ausgeführt werden. Ein Screenshot auf einem statischen Stück Papier wird ihr nicht gerecht. Programme lassen sich jedoch nicht beliebig lange ausführen. Die Computersysteme werden andere geworden sein.
Ein weiteres Problem besteht in der konkreten Bereitstellung von Ressourcen. Dieses Problem ist mit der Komplexität der Bereitstellungs- und Archivierungsinfrastruktur verbunden. Im Rahmen dieser Arbeit wurde mit dem LWMap Service ein Ansatz vorgestellt, der durch klassische Modularisierung funktionale Aspekte voneinander trennt (5.3.6).
Als wichtigster Punkt steht die Trennung der Daten von serverseitig betriebener Hard- und Software. Statische Daten haben den Vorteil, dass sie durch offene Standards wie Unicode in Verbindung mit XML oder RDF allgemein beschreibbar sind und dadurch les- und interpretierbar bleiben. Der Web-Browser als Darstellungselement hat zwar Einschränkungen, ist aber eines der am weitest verbreiteten Werkzeuge und ist durch die millionenfache tägliche Nutzung erprobt. Durch die lange Weiterentwicklung sind Web-Browser stark optimierte Software. Die meisten Web-Browser sind auf den meisten Plattformen verfügbar und basieren auf offenen, gut unterstützen Standards.
Dieser pragmatische Ansatz verspricht dabei auch eine dezentrale und einfache selbstverantwortliche Bereitstellung und Verbreitungsmöglichkeit. Dadurch werden Ergebnisse potenziell von komplexen, wartungsintensiven Strukturen getrennt. Somit sind die Ergebnisse unabhängig von einer komplexen Infrastruktur. Trotzdem folgen die Ergebnisse des LWMap Service dem Linked Data Paradigma, indem auf die Wikipedia verwiesen wird.
Diese Bereitstellung mit Webstandards sichert nebenbei auch den generellen Zugang und das kollaborative Rezipieren. Webstandards sind durch ihren Aufbau und die clientseitige Ausführung des Quelltexts transparent, im Gegensatz zu Bytecode.
Serverseitig sind möglichst freie Technologien zum Einsatz gekommen. In Verbindung mit offenen Daten lassen sich ganze Informationsmaschinen, als virtuelle Maschinen, bereitstellen. Auf diese Weise kann auch die serverseitige Verarbeitungsschicht bereitgestellt werden.

5.5.2 Auswahl und Fokus

Eine Herausforderung für jegliche Graphenanalyse ist die Wahl der Daten aus dem verfügbaren Datensatz und die gewählte Perspektive. Dabei kann eine einfache Auswahl aufgrund von Erhebungen oder konkreten Beobachtungen entstehen. Im Kapitel 3 war der Betrachtungshorizont klar festgelegt. Es wurde die geschlossene Welt von Datenbanken und Einzeldaten behandelt. Die Zusammenführung von Daten ist in diesem Kontext nicht praktikabel, da jede Datenquelle eigene Bezeichnungen und Schlüssel verwendet.
In Kapitel 4 wurde das Paradigma von Linked Data und dem Semantic Web vorgestellt.

Diese Daten folgen dem Paradigma der offenen Welt. Dem Problem der geschlossenen Einzelwelten der Datenbanken wurde hier mit allgemeinen Bezeichnern URIs begegnet. Jedoch wurde auch argumentiert, dass eine Informationsstruktur durch die allgemeine Beschreibung in der offenen Welt sehr komplex sein kann und dadurch Probleme bei der Auszeichnung von Daten und der Verarbeitung entstehen.
Das Problem einer Graphenanalyse steht in einer hypothetischen Linked Data Welt vor der Frage, wie sich Netze bilden und wie weit ein Beobachtungshorizont gewählt werden soll.
Hier kommt es zu mehreren Problemen:
Die Datenmodellierung und die verschiedenen Datenmodelle, die trotz einer Vereinheitlichung vorherrschen. Dazu stellen unvollständige oder unpassende Kategorien weiterhin ein Problem dar.
Durch die Uneindeutigkeit der Werte in dem von DBpedia vorgegebenen Begriffs- und Relationsstrukturen und durch die Open-World-Hypothese sind Datensätze nicht einfach zu verstehen. Beim Generierungsprozess sind wie am Beispiel der Rivalitäten(4.3.2) "Wissensinseln" entstanden.
Das Interessante daran ist die nicht triviale und nicht lineare Entstehungsgeschichte von Wikipedia. In Kapitel 5 wurde das Phänomen weiter untersucht. Wikipedia ist für sich genommen eine offene vernetzte Welt. DBpedia überführt sie auf ein Begriffsystem und Verwendet URIs. Da Wikipedia an sich ist schon eine heterogene Datenquelle ist, stehen die DBpedia-Daten als ein simples einheitliches Modell für Linked Data. DBPedia zeigt sicherlich nicht alle Probleme beim Arbeiten mit Linked Data auf, doch selbst um diese kleine offene Welt zu verstehen, muss ein Ausschnitt gewählt werden.
Abschnitt 2.2.1 zeigte die Grundannahmen für soziologische Betrachtungen, dabei steht ein Einzelner oder eine klar gefasste Gruppe im Vordergrund. Im Rahmen der Wikipedia sind feste Gruppen und Kategorien nur eine kleine Hilfe, da sie unvollständig oder sehr stark in ihrer Granularität schwanken können.
In Abschnitt 2.2.3 wurden anhand von Facebook weitere Vorgehensweisen vorgestellt, wie Auswahlen aus Netzen getroffen werden können. Hier wurde NameGenWeb[29] erwähnt das einen Multi-Ego Ansatz nutzte, es wurde ein Netz aus der Kombination mehrerer Ego-Perspektiven gebildet.
Diese Multi-Ego perspektive spiegelt sich in der Auswahl von Ausgangsartikeln wider. Eine Gruppe von Ausgangspunkten wird gewählt und der Kontext dafür automatisch aus der Datenquelle geholt.
Dieses Vorgehen ist nicht bei allen Datenquellen in der Weise möglich, wie es hier gezeigt wurde. Die hohe Dichte der Wikipedia Page-Links machte dieses spezielle Verfahren jedoch möglich. Erfolgreiche Retrievalverfahren müssen sich auch an der Dichte und der Granularität der Informationen innerhalb der Datenquelle richten. Einer hohen Granularität in Multipartiten Datensätzen kann, wie in Abschnitt 3.6 gezeigt wurde, mit Projektionen entgegengewirkt werden. Dadurch werden die Daten simplifiziert und die Dichte erhöht.

[29] `https://github.com/oxfordinternetinstitute/NameGenWeb` zuletzt gesehen am 03.07.2014

Auswahl entscheidet

Wie in den Auswertungen zur Nutzung des LWMap Service 5.4.3 gezeigt wurde, hat die Auswahl des Fokus einen großen Einfluss auf die Struktur der Ergebnisse. In Abschnitt 5.4.3 ließ sich feststellen, dass über längere Zeiträume einzelne Entitäten immer wieder für das Retrieval genutzt wurden.
Das widerauftreten von Entitäten wurde in den Auswertungen im Abschnitt 5.4.4 genauer untersucht. Über die In-Degree-Verteilung wurde sichtbar, dass die für die Suche gewählten Knoten sich sehr stark im eingehenden Grad unterschieden. Es wurde davon ausgegangen das ein Artikel mit hohem eingehenden Grad bekanntere Dinge abbilden. Die gesuchten Knoten hatten hier einen höheren eingehenden Grad als die in den Visualisierungen fokussierten und geöffneten Knoten. Artikel, die für die Erstellung der Visualisierung genutzt wurden, sind also bekannter und von generellerer Natur als die anderen untersuchten Knoten.
Auch wurde gezeigt, dass sich die Ergebnisse der intendierten Suchen sehr stark von zufälligen zusammengestellten Suchbegriffen unterscheiden. Anhand der Zufallszusammenstellungen konnte gezeigt werden, dass der Recall bei den intentionalen Kombinationen (den Suchen der Nutzer) sehr viel höher war. Die großen Unterschiede in den Mengen der Kontextknoten zeigen das sehr deutlich. Es ist davon auszugehen, dass Vorwissen für ein erfolgreiches Retrieval und damit für einen sinnvollen Ausschnitt nötig ist. Der DBPedia-Pagelink Datensatz ist nicht so dicht, dass zufällige Suchen eine sinnvolle Ergebnismenge liefern.

Dominante zentrale Knoten

In vielen Quellen gibt es Hub-Knoten, die einen hohen Grad und eine hohe Zentralität Aufweisen und damit Netze dominieren. Diese offensichtlichen Knoten sind dabei nicht sehr informationsträchtig und stehen der Verarbeitung teilweise im Weg.
In extremen Fällen, in denen fast alle Knoten mit einem einzigen Hub-Knoten verbunden sind, könnte dieser eher als Überschrift dienen denn als Knoten.
Im Rahmen dieser Arbeit kann Macbeth (3.2.1) als Beispiel genannt werden. Doch auch andere Prosatexte mit einem Hauptcharakter weise ähnliche Eigenschaften auf, wenn sie als Personennetz betrachtet werden.[86] Im speziellen dürfte das für Geschichten mit Ich-Erzähler gelten, dieser sollte mit allem verbunden sein, da er per Definition immer und überall dabei ist.
In anderen Teilen könnten sie komplett vernachlässigt werden wie bei den Visualisierungen aus der Wikipedia. Hier sind vor allem Artikel wie der über die Vereinigten Staaten von Amerika zu nennen.[23] Diese sind so stark verlinkt, dass sie fast immer vorkommten. Diese Artikel umfassen jedoch sehr viel mehr und würden daher nicht als Überschrift infrage kommen.
Im LWMap-Service wurde diesem Phänomen begegnet indem für den Ausschnitt wichtige Knoten, wenn alle Links des Artikels auch in der Visualisierung gezeigt wurden, groß dargestellt wurden. Wenn die Bedeutung, die Anzahl der Links, jedoch weit über den be-

trachteten Ausschnitt hinausgeht, wurden diese generell im DBPedia Datensatz zentralen Knoten klein dargestellt.

Filter, Projektion und Komponenten

Die Anzahl von Knoten und Kanten sind entscheidende Werte, wenn eine lesbare Visualisierung angefertigt werden soll. Dafür Offensichtliches gefiltert werden. Der vorgestellte Ansatz der Projektion ist dabei eine Möglichkeit in multipartiten Netzen die Menge der Knoten zu verringern, wie an der Arachne Datenbank gezeigt (3.5). Der Nachteil ist jedoch, dass die Menge der Kanten explodieren kann und sehr dichte Komponenten entstehen können.
Wenn in diesem Fall gefiltert wird, dann muss dargelegt werden, welche Relationen warum gefiltert werden (3.6.5). Durch die Projektion bleibt der Zusammenhang von Komponenten wie im Ausgangsdatensatz erhalten. In der Praxis könnte durch Projektion verhindert werden, dass Graphen durch Filter auseinanderfallen. Dafür könnte Ad-hoc ein bipartites Netz aus den zu filternden Knoten und dem Rest erstellt werden.
Dieses Problem tritt auf, wenn man die verbindenden Knoten und ihre Kanten nur filtert. Eine Projektion erhält im Gegensatz zum Filter den Zusammenhang bzw. die Komponenten. Bei Arbeiten mit kräftebasierten kann auf diese Weise das Abdriften von Knoten verhindert werden, da der Gesamtzusammenhang nicht verloren geht.

Dekonstruktion komplexer Netze

Große und komplexe Datenstrukturen sind nicht in jedem Fall dafür geeignet, komplett als Netz abgebildet und verarbeitet zu werden. Um diesem Problem entgegen zu treten, wurde das Meta-Matrix-Verfahren verwendet (Abschnitt 3.6). Hierbei wurde der multipartite Gesamtgraph in bipartite Graphen umgewandelt. Dabei kommen Komponenten zum Vorschein die sich um bei bestimmten Realtionstypen bilden. Durch anschließende Projektion konnten durch die Analyse der Maßzahlen interessante Netzkomponenten aufzeigen. Die meisten Zusammenstellungen lieferten jedoch nur fragmentierte Netze oder Netze, die aus komplett verbunden Kleinstkomponenten bestanden. Die Projektion konnte bei der Feststellung helfen, ob sich transitive Strukturen bilden, die potenziell interessant sind und nicht nur Zuordnungen die sich nicht überschneiden, sind.
Relationale Datenbanken haben durch ihre Konzeption eine klare Klassenstruktur. In Begriffssystemen wie Wikipedia lassen sich Datensätze bzw. Artikel nicht einfach eineindeutig klassifizieren. Dabei kommt es auf die bestehende Klassifikation an. Bei sich überlappenden, nicht diskriminierenden Klassen können keine bipartiten Teilnetze ermittelt werden. Gerade im Bereich des Semantic Web kann es problematisch sein, eine sich ausschließende Parität zu bestimmen, da Entitäten bewusst mehrfach klassifiziert werden können. Das DBPedia Beispiel war in dieser Hinsicht interessant, da es eine Menge von Kategorien gibt, die überlappend und unvollständig sind.
Die Betrachtung einzelner homogener Cluster kann nach der klassischen unipartiten Lesart interessant sein. Eine weitere Untersuchung könnte ermitteln, welche Kombinationen von Prädikaten weitreichende Netze bilden. Dies könnte ähnlich realisiert werden wie bei

der Untersuchung der Arachne-Datenbank. Ein solcher kombinatorischer Ansatz wäre rein maschinell implementierbar.

5.5.3 Qualität und Vollständigkeit

Eine Netzdarstellung gibt Informationen über einen Datenbestand. Ob damit eine bestimmte Fragestellung beantwortet werden kann, muss ermittelt werden. Unvollständige Daten, mit Bias erfasste Daten oder einfach unerwartet codierte Daten können ein verzerrtes Bild liefern. Bei der Verarbeitung von netz basierten Daten müssen daher typische Verzerrungen erkannt werden. Nur so können Daten zur weiteren Verarbeitung aufbereitet werden. Die Wahl von Abfragen und Filtern ist wichtig um eine oder mehrere Datenquellen auf eine Anwendung oder eine Fragestellung hin nutzbar zu machen. Die Verzerrung durch Unvollständigkeit kann auch zur Evaluation des Stands eines Datenbestands dienen. Die meisten Datenbanksysteme sind darauf ausgelegt zu wachsen. Oft ist es schwer, von einem vollständigen oder endgültigen Zustand zu sprechen. Beim Aufbau und kontinuierlicher Weiterentwicklung von Datenbeständen können Lücken gefunden und dann gezielt geschlossen werden. Die Netzstruktur kann zu einem Kriterium für Vollständigkeit werden, wenn bekannt ist, welche Struktur die Daten bilden müssten.

Unerwünschte Fragmente und Überlagerungen

In den Untersuchungen in dieser Arbeit ließen sich immer wieder unerwünschte Fragmente finden. Durch die visuelle Analyse und anhand von Maßzahlen ließen sich diese Fragmente identifizieren.
Hier ist bei Macbeth (Abschnitt 3.2.1) die Rolle “mac-all” zu nennen. Diese “Rolle” kennzeichnet Text, den alle Charaktere auf der Bühne sprechen und bezeichnet nicht mal eine konstante Gruppe.
In der Betrachtung von Bauwerken auf Buchseiten in der Arachne Datenbank (Abschnitt 3.6.3) ergaben sich Fragmente aus der Granularität der Daten. Buchseiten zeigen oft voneinander unabhängige Darstellungen, die keinen geografischen Bezug haben. Das ist kein Fehler, sondern nur ein Problem, wenn die Verbindungen als geografische Nähe interpretiert werden.
Eine Überlagerung die aus der Retrievalmethode heraus entstand in der Analyse von Objekten und Sammlungen in der Arachne Datenbank (Abschnitt 3.6.4). Durch die Projektion wurden dabei zwischen allen Sammlungen Verbindungen erstellt, denen die gleichen Objekte zugeordnet waren. Für eine Betrachtung von direkten Verbindungen zwischen Sammlungen und wurden so Informationen überlagert. Eine Lösung dafür wird in Abschnitt 5.6.2 besprochen.
Ein weiteres interessantes Fragment fand sich in den Betrachtungen der Rivalitäten in der DBPedia (Abschnitt 4.3.2). Hier wurde die fiktive Motorrad Gang “Sons of Anarchy” in eine Komponente realexistierender krimineller Organisationen durch das Retrieval erfasst. Dieser Fehler ist interessant, da er eine Entgrenzung des Realen darstellt. Fiktive Information wird mit realer Information gemischt.

Maßzahlen

Die Verteilung der Gradzahlen können einfache Hinweise auf Probleme bei Matchings liefern und bei der Optimierung von Vorgängen helfen, wie in Abschnitt 3.2.2 beschrieben.
In der Analyse der Arachne in Abschnitt 3.6.2 wurden hauptsächlich Maßzahlen zur Identifikation von interessanten Netzen verwendet. Hier ließen sich Überlagerungen durch die Dichte der aus dem Projektionsverfahren hervorgehenden Daten erahnen.
Bei Maßzahlen wie den Zentralitäten ist die Vollständigkeit manchmal ein kritischer Faktor und beeinflusst die Ausprägung signifikant.[61]
Wie genau unvollständiges Wissen sich auf Maßzahlen auswirkt, ist dabei auch von der Art des fehlenden Wissens und den Wissensbeständen abhängig.[145]
Daher muss bei tiefer gehenden Betrachtungen immer der Hintergrund der Datenerhebung betrachtet werden. Hier ist ein Verständnis der Maßzahlen unumgänglich. Durch die in diesem Zusammenhang oft verwendete Betweenness Centrality kann beispielsweise durch das Fehlen einzelner Brückenverbindungen stark beeinflusst werden.
Die Verwendung von Zentralitätsmaßen ist aber bei der explorativen Analyse mit einem gewissen Vorwissen eines Datenbestandes uninteressant, da dem Experten die zentralsten Entitäten des Datenbestands bekannt sind. Einem Fachmann für Macbeth und wahrscheinlich auch jedem Laien wird klar sein, dass Macbeth die zentrale Person im gleichnamigen Stück ist. Jedoch hilft es dem Experten, durch unvorhergesehene Verschiebungen dieser Werte systematische Lücken und Fehler im Datenbestand zu erkennen.
Ein anderer Ansatz, in dem diese Werte wieder einen Sinn machen, ist das Vergleichen von Versionen des gleichen Stücks. Hier können so schnell Verschiebungen des Fokus des Stücks in Hinsicht auf einzelne Charaktere ermittelt werden.

Orte

Bei den Betrachtungen von Zusammenhang müssen Ortssysteme, die in Datensätzen oft vorhanden sind, besonders betrachtet werden.
Orte haben beispielsweise in der Wikipedia in vielen Bewertungsmetriken eine hohe Relevanz.[23] Sind alle möglichen Interaktionen und Überlappungen vorhanden, bildet sich an Orten zwangsläufig ein sehr zentrales Netz.
Gerade die Einfachheit und Klarheit von Orten und der Ortszugehörigkeit macht eine umfassende Erfassung von Orten sehr einfach. Da in Wikipedia auch Bots Artikel bearbeiten, kann angenommen werden, dass hier gute automatische, wie allgemeingültig formulierbare Auszeichnung erfolgen kann.
Durch die klar definierbare geografische Welt kann meist widerspruchsfrei belegt werden, dass ein Link angebracht ist und eine gewisse Relation widerspiegelt. Die Überprüfung ist von fast jedem anhand von Karten möglich. Diese Eigenschaft wurde in Abschnitt 3.3 beispielsweise genutzt, um eine Abstimmungsabweichung zugänglich zu machen. Durch eine lange Tradition der Vermessungstechnik ist dieses Wissen meistens unumstrittenen. Diese Gewissheit mag abnehmen, wenn man den Gesamtkontext aller Zeiten nimmt, vor allem die genaue Abbildung von Transaktionsstadien. Doch wird dieses Wissen wahrscheinlich

vollständiger und gewisser sein als die meisten anderen Informationen aus historischem Kontext.
In einer Community wie Wikipedia ist ein Link auf eine geografische Information daher recht widerspruchsfrei und wird eher keine Diskussion unter den Nutzern auslösen. Anders sieht es mit anderen Informationen aus wie beispielsweise künstlerisch abstrakte Beziehungen zwischen historischen Personen, Glaubenssystemen und gesellschaftlichen Entwicklungen. Hier wären die Philosophenrelationen aus Wikipedia (4.3.1) zu nennen.[153]
In dieser Arbeit hat sich die Dominanz von Ortsbezügen in der zufälligen Rekombination der Suchanfragen an den LWMap Service gezeigt (5.4.4). Für tiefer gehende oder konzeptionelle Information ist es daher fraglich, inwiefern gerade Ortsinformationen redundant oder zu weitläufig sein können. Eine Verbindung zu einem Staat ist fast immer eine sehr weitreichende Beziehung.
Ortsinformationen sind in komplexen Datensätzen, die multipartite Netze bilden meist vorhanden und potenziell sehr vollständig. Wenn Beziehungen von Orten untereinander abgebildet sind, dann bilden sie eine einzige Komponente und liegen wahrscheinlich auf vielen kürzesten Pfaden zwischen Entitäten, die keine Orte sind. Durch diese Eigenschaften können sie für den Betrachter interessante Informationen überdecken und Verbindungen herstellen, die eigentlich nicht informativ sind.

5.5.4 Erweiterter Zugang

Der erweiterte Zugang zu Datenbeständen ist ein wichtiges Thema in einer Welt, in der bei fast allen Vorgängen mehr oder weniger automatisch Daten entstehen. Als Beispiel kann hier der LWMap Service gezeigt werden, durch dessen Verwendung Daten erzeugt wurden.
Die hier vorgestellten Systeme haben natürlich das Problem, dass nicht alle Datenbestände vollständig und verlustfrei dargestellt werden können. Das ist jedoch auch nicht immer von Interesse.
Die Möglichkeit der Fokussierung und der Reduktion von Komplexität ist wichtig, damit große Datenbestände erfahrbar, aber auch vergleichbar und zugänglich werden.
Zugang kann auch auf verschiedenen Stufen erfolgen, wie die Beispiele aus frühen Computerspielen (1.2) aber auch an den Ansichten der Arachnedatenbank (3.6.2) und den Wikipedia Visualisierungen (5.2.2) gezeigt wurde. Die Verbindung von Netzen mit Kategoriesystemen wird dabei voraussichtlich von besonderem Interesse sein. Die Multiparität und Ungewissheit in Datensätzen ist dabei ein Fakt, dem vor allem im LWMap Service Rechnung getragen wurde. Wie konkrete Erweiterungen von Interfacetechnologien aussehen können, ist in Abschnitt 5.6.1 beschrieben.

Serendipität

Etwas das sich aus Browsing ergeben kann, ist Seredipidtät. Ein Beispiel für Serendipität ist das Auffinden von interessanter Literatur beim Stöbern in einem Bücherregal im Gegensatz zur Suche nach einem Buch. Dieses Beispiel wird als Argument gegen digitale Textsammlungen verwendet, da das Stöbern nicht mehr möglich ist.[251] Dabei ist dieses

scheinbar zufällige Auffinden natürlich nicht wirklich zufällig, da man sich in einem sortierten Raum befindet, einer Bibliothek.
Die klare Ordnung muss dabei nicht ersichtlich sein und ist auch nicht für alle Bibliotheken gleich, doch wird es ein wie auch immer geartetes Ordnungskriterium geben. Daher kann es auch sinnvoll sein, in verschiedenen Bibliotheken zu lustwandeln, da sich der Inhalt und die Ordnungskriterien zwischen Bibliotheken unterscheiden. Als digitale Alternative wurde in dieser Arbeit ein Visualisierungskonzept, das eine eigene Raumordnung aufstellt, vorgeschlagen.
Verstärkend könnte hier die Longtail-Verteilung von Popularität sein. Beliebtes wird beliebter, daher präsenter dargestellt, bis alles andere verdrängt wird. Ein Fortkommen durch das Auffinden von Neuem wäre so nicht möglich.
Dieses Prinzip ist dabei im Multipartiten und unfokussierten Kontext wie Wikipedia interessant da hier viele Einflussfaktoren aufeinandertreffen können und nicht nur ein Klassifizierungsschema besteht, sondern eine ganze Menge von Einzelklassifikationen und Autoren zu den Informationen beitragen. Diese Bedingungen scheinen bei der serendipitären Vorgehensweise besonders vielversprechend.[95]
In dieser Arbeit wurde das Auffinden von Neuem im LWMap Service versucht. Dabei wurde eine netzbasierte Gewichtung vorgeschlagen (5.3.4). Diese bezog sich auf die lokale Wichtigkeit im Hinblick auf den abgebildeten Kontext.
Weiterhin wurde der Hidden Context untersucht (5.3.6). In der Analyse ist diese Art des Filterns am häufigsten vorgekommen. Der Hidden Context zeigt die passiv verlinkten Artikel. Beim einfachen Browsen sind diese Artikel nicht erfassbar.

Visualisierung für die Spezialisten?

Die hier vorgestellten Visualisierungen haben den Zugang zu Daten erleichtern können, das jedoch nicht beliebig intuitiv. In der Frage von Userinterfaces und Bedienung sind “intuitiv” und “natürlich” schwierige Begriffe. Es gibt natürlich Aussagen über das Sinnsystem die empirisch belegt sind. Jedoch bringt jeder Mensch Vorwissen über Systeme mit. Daher können diese Bedienparadigmen zu einer Zeit funktionieren und zu anderen nicht. Die Verwendung von Touchscreens war vor der Erfindung des Smartphones nicht sehr verbreitet.
Im Zusammenhang kann hier an den Begriff der digital Natives[202] angeknüpft werden. Vereinfacht steht der Begriff für die Annahme, dass Technologie die Menschen, die mit ihr aufwachsen, in ihren Begabungen und ihrem Lernen verändert, sie können intuitiv mit Technologie umgehen.[24]
Ein weniger akademisches Argument für das Anwachsen von Kompetenz wären die Verkaufszahlen von Videospielen. Denn die in Videospielen vorhandenen Konzepte finden sich auch in dynamischen Visualisierungen wieder. Nebenbei könnte spielerisch diese Art des Zugangs trainiert werden. Der Vorteil bei der Nutzung von Spielen liegt in deren intrinsischer Motivation. Ansichten, die in Spielen, von einer breiten Masse an Menschen verwendet werden, können auch außerhalb des Kontexts von Spielen verstanden und bedient werden. Das Problem an Spielen ist, dass die Motivation durch den Spieler gegeben

sein muss. Ein Spiel, das keinen Spaß macht, wird nicht gespielt und kann nichts vermitteln.
Es bleibt die Frage, inwiefern sich einzelne Ansichten oder dynamische Systeme für die breite Masse an Nutzern etablieren können. Ein Potenzial für die Verwendung von Netzen wäre eine Abkehr vom linearen Paradigma der heutigen Ergebnisansichten in Form von Listen. Dabei werden wahrscheinlich die digital Natives weniger Probleme haben, da sie durch den frühen Umgang mit visuellen und interaktiven Medien, komplexere Interfaces bedienen können und weniger Probleme mit der Adaption von komplexen Systemen haben. Daher können aus komplexen Nischentechniken zukünftig Methoden entwickelt werden, die auch für ein größeres Publikum nutzbar sind.
Eines der komplexesten Themen ist die Entstehung von Bildern und Visualisierungen. Eine Darstellung, deren Hintergrund und Erstellprinzipien man nicht kennt und die trotzdem einen Anspruch auf Wahrheit reklamiert, ist potenziell suspekt. Das ist nicht allein auf das Misstrauen gegen den Erstellenden zurückgeführt. Da Visualisierungen Aussagen untermauern sollen, müssen sie verstanden werden.
Ob sich Kompetenz für diese Themen in der breiten Masse der digital Natives wiederfinden wird, ist nicht klar, da viel Komplexität hinter guten Nutzeroberflächen und komplexen Applikationen versteckt bleibt.[30]

5.6 Ausblick

In diesem Abschnitt werden nun die weiteren Möglichkeiten der Datenverarbeitung auf Grundlage der Ideen aus dieser Arbeit besprochen.

5.6.1 Erweiterte Interfaceelemente

Interaktive Darstellungen von Netzen können helfen, bestehende Konzepte von Interfaces und Daten zu erweitern. Die im Folgenden beschriebenen Technologien können aus den Konzepten in dieser Arbeit hervorgehen.

Facetted Browsing und Netze

Ein Ansatz, der dem Problem der unvollständigen Daten, wie es in dieser Arbeit öfters verwendet wird, sehr erfolgreich entgegentritt, ist das Facetted Browsing. Ein Ansatz für dieses Vorgehen wurde in LWMap-Service mit dem Filtern über die Kategorien als Facetten in der Darstellungsoberfläche realisiert.
Facettedbrowsing stellt eine Menge von Features zum Filtern bereit, die mit mehrfach Kategoriensystemen umgehen können. Dadurch können effizient Eingrenzung eines Kontexts geschaffen werden. Die Abbildung als Netz könnte hier unzusammenhängende Listen abbilden und ergänzen. Überlappenden Kategorien kann so mehr Ausdruck verliehen werden

[30] Hier wird auch von digital Naives (ohne t) gesprochen, da zwar früh und umfassend mit digitalen Technologien umgegangen wird, der Hintergrund und die Funktionsweise jedoch versteckt bleiben. Eine Blackbox entsteht, deren Innenleben nicht verstanden wird.

als durch Listen. Dadurch würden Beziehungen und Vorkommensüberschneidungen darstellbar, die vorher so nicht wahrnehmbar waren. Auch sind durch die Einschränkungen Datencluster einfacher zu identifizieren.
Das wäre beispielsweise mit Hilfe von Projektionen möglich. Der Vorteil dabei ist, dass die Kanten kumuliert und gewichtet werden können. Die Menge der Knoten nimmt dabei ab, was es einfacher macht einzelne Knoten zu identifizieren. Das Aufprojizieren mehrerer Kategorien könnte so den Zugang zu Daten erleichtern und einen Fokus zulassen.

Text++

Die strukturierenden Eigenschaften von Texten auf semantischem Niveau können weiteren Zugang zu Texten geben. Hier könnten alternative Ansichten helfen. Dabei stehen weniger die genauen Ansichten im Mittelpunkt sondern die Übersichten. Dabei bietet Text eine der besten Quellen für erfassbares Wissen. Dieses Potenzial ist dabei auch in komplexen literarischen Werken vorhanden.
Bühnenstücke (3.2.1), Rechtsliteratur (3.2.2), Kataloge, Wörterbücher und sind durch ihren klaren Aufbau einfach zu analysieren. Die in 3.2.1 gezeigten Visualisierungen mit Textbeschreibungen ist dabei ein erster Zugangspunkt. Einem Experten wird dieser Zugang nicht viel bringen. Einem völlig Unbedarften werden die Worte aber wenigstens einen Anhaltspunkt geben, was es wo gibt. Daher wäre hier wirklich eine explorative Übersicht gemeint.
Mit HTML gibt es eine Darstellungsvariante, die potenziell für alle Textformen möglich ist. Durch die Ähnlichkeit zu XML ist damit ein Potenzial für Zweitauswertungen gegeben, da sich die Ansteuerung nicht groß von der Ansteuerung von XML unterscheidet.
Für die automatische Visualisierung von diesen nach Standards ausgezeichneten Texten, wie TEI, ist es interessant, “Backrezepte” zu erstellen, die sich an die Menge der Daten anpassen lassen.
Die Bereitstellung von Werkzeugen anhand von Datenstandards hätte dabei einen sich selbst verstärkenden Effekt. Standards würden öfter genutzt, da ein Mehrwert aus ihrer Nutzung entsteht.
Annotierter Text bekäme eine Art automatisch erstellbarer Außenansicht. Diese ließe sich mit einem Inhaltshalts- oder Stichwortverzeichnis vergleichen, nur dass es nicht auf Text, sondern einer Visualisierung beruht.
Eine Wiki ist in diesem Zusammenhang eigentlich nur ein weiterer Schritt in eine größere Welt. Diese besteht nicht aus einem, sondern einer verbundenen Menge von annotierten Texten. Gerade im Zusammenhang mit E-Book-Endgeräten wäre so etwas interessant.

Erweiterte Zoom Analogie

Im Umgang mit Filtern wurde der Vorschlag gemacht, Filter durch Projektionen zu ersetzt. Das soll eine Verbundenheit eines Graphen zu gewährleisten. Knoten würden gefiltert und mit in die Kanten einfließen. Eine andere Analogie bei Karten ist das Hereinzoomen. Höhere Zoomstufen, die nur wenige Knoten mit wenigen Kanten abbilden, könnten mit parallelen Erklärungsmustern für Kanten hinterlegt werden. Dabei würden die Kanten

durch zwei Kanten und einen Knoten ersetzt, wie beim RelFinder, nur dass der Knoten nicht den Relationstypen aufzeigen würde. Die Quellen der aufprojizierten Kanten würden stattdessen gezeigt. Sollte der Graph nicht aus aufprojizierten Kanten bestehen, könnten Erklärungen für eine Kante gesucht werden. Die Kante würde durch etwas ersetzt, was die Kante erklären könnte.
Im Beispiel die Projektion des Graphen aus Macbeth(3.2.1) würden die Szenen durch Charaktere als Kanten verbunden. Bei hohen Zoomstufen könnten die Kanten in die Akteure zurückübersetzt werden. Eine Auflösung ohne Hintergrund wäre es, nach überlappenden Worten in den Szenen zu suchen und diese zwischen den in der Zoomstufe sichtbaren Szenen darzustellen.
Das würde auf Knotenebene verhindern, dass nicht zu viele Knoten dargestellt werden müssen. Dabei wäre natürlich eine Darstellung für diese spezielle Zoomstufe notwendig, da die hinzukommenden Knoten lokal positioniert werden würden. Durch die Kombination aller sichtbaren Szenen auf einer Zoomstufe ließe sich ein inhaltlich wie lokaler Filter erstellen lassen.

5.6.2 Erweiterte Untersuchungen und Nutzung

Im Rahmen dieser Arbeit wurden eine Menge von Datensätzen auf ihre Eigenschaften als Netze untersucht. Dabei wurden keine erschöpfenden Analysen angefertigt, sondern es wurden einige Anfangsbetrachtungen gemacht. Da diese Arbeit eigenständig verfasst wurde, war zu vielen Themen auch kein Expertenwissen vorhanden. Wissenschaftliche Projekte, in denen Software entwickelt wird, kommen jedoch oft nicht ohne Experten aus verschiedenen Fachgebieten aus. Die hier vorgestellten Analysen sind daher natürlich nicht erschöpfend.
In dieser Arbeit ging es um den Prozess der Erkenntnis, wie textliche und geisteswissenschaftliche Quellen aufgebaut sind und ob sie Potenzial für die Betrachtung und Verarbeitung als Netz haben. Dabei wurde versucht, von der ideellen Welt, wie sie in Datenmodellen, Datenbanken, Auszeichnungsstandard und Ontologien beschrieben, wird Abstand zu nehmen und die reale Welt der expliziten Daten zu betrachten. Denn erst dieser Schritt macht es möglich, das Potenzial von Datenquellen für die weitere Verarbeitung zu evaluieren.
Generell sind in dieser Arbeit einige generalisierbare Ansätze der Extraktion vorgestellt und auch generalisierbar implementiert worden. Diese Ansätze könnten ausgebaut und mit den vorgestellten Mitteln ohne Weiteres online zugänglich gemacht werden.
Im Folgenden werden weitere Ansätze zur Analyse und Verwendung hervorgehoben, die aus den Workflows und Daten in dieser Arbeit hervorgehen.

Das Sammlungstransaktionsnetz

Projektionen haben, wie am Arachnedatensatz untersucht, den Vorteil, dass sie sehr dichte Netzwerke erzeugen können. Dabei sind nach dem explorativen Auffinden Nachbearbeitungen nötig. Dabei können einfache Eigenschaften helfen Analysen auszuführen.
In Abschnitt 3.6.4 wurden die in der Arachne hinterlegten Sammlungen mit den ihnen

zugehörigen Objekten gezeigt. Da Objekte nur temporär in einer Sammlung sind, ist es möglich, einen Fluss durch das Netz aus Sammlungen zu beschreiben. Dabei kann über die Zeit des Bestehens von Sammlungen zwischen Gründung und Auflösung geschlossen werden, welche Abfolge die Stücke nahmen. Diese Information ist der Arachne-Datenbank nicht systematisch erfasst worden, sie ist jedoch menschenlesbar enthalten.
Es müsste ein Regelsystem aufgestellt werden, das nur theoretisch mögliche Kanten zulässt. Die direkte Transaktion von Objekten zwischen Sammlungen, deren Existenz sich nicht überschneidet oder wenigstens aneinandergrenzt, wäre damit ausgeschlossen.
Möglich wäre auch eine Filterung durch ein Transaktionsspielfeld. Zur Erstellung dieses Spielfeldnetzes würden alle zeitlich möglichen Transaktionen zwischen den Sammlungen errechnet. Eine Kante würde zwischen allen Sammlungen angelegt, die zeitlich aneinandergrenzen. Hierzu gehören Überschneidungen in der zeitlichen Existenz sowie kurze Zeiträume zwischen Ende einer Sammlungen und der Gründung einer neuen Sammlung.
Die Überbrückung der Lücken kann verschieden begründet werden, so sind Transporte, ungenaue Datumsangaben, die Zeit von Versteigerungen und zuletzt kleinere menschliche Fehler bei der Eingabe in die Datenbank zu nennen. Fehlende Daten könnten beispielsweise durch Textanalysen zum Vorschein kommen. Gerade Informationen zu Auflösungen können weiterhelfen, da hier Versteigerungen und andere Sammlungen textlich gesammelt wurden. Des Weiteren könnte man eine Existenz Zeitspanne in allen erwähnten Sammlungen durch Auflösung und Entstehen finden. Wenn diese nicht vorhanden sind, kann dies durch die Lebenszeit der Sammler errechnet werden.
Die Schnittmenge aus Kanten des Graphen aus der einfachen Projektion und dem im vorhergehenden Absatz vorgestellten Netz würde dann ein genaueres Modell ergeben. Dieses Regelsystem könnte auch als Berechnungsfunktion bei der Projektion hinterlegt werden (3.5.1). Dabei würden keine Kanten erzeugt, wenn Sammlungen nicht zeitlich aneinandergrenzen, daher müsste es einen Zwischenbesitzer geben. Dieser Zwischenbesitzer ist im Zweifelsfall nicht in den Daten vorhanden, hierfür könnten beispielsweise Blanknodes eingefügt werden, dieses Vorgehen wird bei RDF verwendet, um unbekannte oder unbezeichnete Entitäten zu beschreiben. Dabei wird keine Aussage über die Granularität gemacht. Der Blanknode steht für eine Ungewissheit. Wenn trotzdem noch Lücken existierten, müssten sie durch “neue” Blanknodes ersetzt werden. Eine solche Art von bekannten Unbekannten könnte dabei natürlich auch interessant sein. Denn eine sehr schwer zu lösende Aufgabe kann erfolgsversprechender sein als die effizienteste Lösung aller möglichen Optionen.
Der Fluss und die Transaktionen zwischen Sammlungen und Museen könnten einen mehr oder weniger gerichteten azyklischen Graphen bilden. Das würde die Visualisierung um einiges effektiver machen, da eine Zeitachse als ordnendes Element eingeführt werden könnte. Bei Kulturgütern ist es aus wissenschaftlicher Sicht interessant nicht nur das Objekt zu besitzen, sondern auch den Kontext zu kennen zu kennen, aus dem es stammt. Durch Methoden des illegalen Kunsthandels, der bewusst Provinienzen verschleiert, wird diese Art relevanter Information nicht nur unterdrückt, sondern auch verzerrt. Daher könnte eine solche Karte vielleicht helfen, die Plausibilität von Provinienzangaben zu prüfen.

Verwertung: Simulationen und Spiele

Simulationen beispielsweise von agentenbasierten Systemen und im Spiele können eine Möglichkeit sein, die aus diesen Verfahren extrahierten Daten zu nutzen. Simulationen können beispielsweise helfen, die Konstruktion von Netzen anhand von in Software implementierten theoretischen Modellen zu erforschen. Dabei wäre das extrahierte Netz eine Zielkonstellation, die es nachzubilden gälte. Auf der anderen Seite kann ein Netz als Spielfeld dienen, auf dem Transaktionen ausgeführt werden.
Eine praktische Anwendungen findet man für die extrahierte Struktur das Netz aus Bandenrivalitäten, in Abschnitt 4.3.2 beschreiben. So könnte aus der Allianz und Rivalitätsstruktur der Startpunkt einer Agent-Based-Simulation für Konflikte hervorgehen. Auch wäre es möglich, dieses Netz als Starkonfiguration eines Computerspiels zu nutzen.
Der große Vorteil dieses Grundgraphen aus dem Semantic Web und Linked Data ist, dass je nach Erweiterung einer solchen Simulation weitere Faktoren gesucht und aus den Quellen hinzufügt werden können, z.B. gemeinsame Geschäftsfelder, geografische Überschneidungen oder die ethnische Struktur der Organisation. Diese Faktoren sind durch die Erweiterung der vorgestellten SPARQL-Befehle realisierbar.
Die Daten aus den Shakespeare-TEI-Dokumenten könnten die Grundlage für ein Serious-Game bilden. Ein Projekt in diesem Rahmen ist "Will in Town", das im Rahmen einer der Vortragsreihe "A Party For Will" von Studierenden der Uni Köln erstellt wurde.[265, 223] "Will in Town" ist eine Smartphone App. Das Spiel ist dadurch geografisch interaktiv. Es gibt geografisch lokalisierte Quests an Orten über die Kölner Innenstadt verteilt. Thematisch wird Shakespeares Werk in den Quests mit Minispielen, beispielsweise ein Quiz zu den Charakteren, genutzt. Im Vergleich könnten die Verarbeitungen und Visualisierungen nur ein Teilaspekt eines recht aufwendigen Stücks Software sein. Zur weiteren Recherche könnten in verschiedenste Online-Quellen[31] zum Close Reading durch einfache Links referenziert werden.

Kleine "fiktionale" Welten

Neben der realen, allumfassenden Welt wie sie in Wikipedia versucht wird abzubilden, gibt es eine Menge von fiktionalen Welten. Diese Welten bestehen für sich und sie sind komplett erfassbar. Dabei ist natürlich eine Interpretation möglich, doch sind erdachte Welten wie Mittelerde[249], das Star Wars Universum[268], Comic Universen zum einen per Definition erfasst, da sie durch ihre Werke bestehen. Hierbei stellen gut gepflegte Wikis von Fans für viele solcher Universen Metainformation bereit, die auch sehr komplett sind.
Sie bilden eine Metadatenstruktur um fiktionale Räume. Diese Räume, da sie auch schon lange existieren, haben eine enorme Komplexität aufgebaut.
Für Autoren besteht hier die Herausforderung neue Werke zu verfassen, ohne Widersprüche in der Welt zu erzeugen. Das ist ein Problem für fast alle langjährige episodenhafte

[31] Kandidaten wären beispielsweise eine Eigenentwicklung, das Portal des Verlags JStore zu Shakespear[254], da hier auch sehr viele Sekundärquellen eingebunden sind, oder die Beispiel Web-App der eXist Datenbank [32] in der über die IDs Dokumentstellen eingebunden werden können.

Veröffentlichungen mit zusammenhängender Geschichte.
Exemplarisch steht hier das Star Wars Universum. Nach der Veröffentlichung der Hauptfilme wurden viele Zweitverwertungen lizenziert, wie Comics, Computerspiele, Fernsehserien. Diese spielten im erweiterten Star Wars Universum. Der offizielle Kanon wurde nach dem siebten Kinofilm auf die Filme, sowie die Serien Clone Wars beschränkt.[33]
An dieser Stelle können netzbasierende Informationen für Autoren interessant sein, um mögliche Überschneidungen von Situationen, Personen oder auch Gruppen von Personen und ihre Beziehungen abzubilden.
Auf der anderen Seite könnten auch den Fans dieser Welten solche Werkzeuge nutzen, um weiter in diese Welten einzutauchen. Hier könnten auch ausgefeilte Filterkonstrukte helfen, keine interessanten Fakten zu verraten (Spoiler). Auch kann es für Fans Interessant sein, inwiefern sich literarische Vorlage und filmische Umsetzung unterscheiden. Hier wäre die Fantasy-Romanreihe "A Song of Ice and Fire"[3]und deren filmische Umsetzung in der Serie "Game of Thrones"[102] zu nennen.
In diesem Kontext ist eine Entgrenzung interessant, da die wirkliche Welt ohne fiktionale Welten bestehen kann. Fiktionale Welten funktionieren jedoch nicht ohne Bezug zur realen Welt. Hier treten banale real existierende Phänomene wie Zeit, Schwerkraft, intelligentes Leben (Menschen), Handlungen und Beziehung auf. Die Frage nach sinnvollen Grenzen kann da nur mit der Gegenfrage nach dem Anwendungszweck beantwortet werden. Auch lehnen sich fiktionale Welten oft an Ereignisse in der realen Welt an. Auch das stellt eine Vernetzung dar.
Im pädagogischen Sinne ließen sich solche Verbindungen nutzen, um Medieninhalte aus der Freizeit mit realen Ereignissen und Phänomenen ins Verhältnis zu setzen.
Ganz konkret ließe sich dies mit Wikis realisieren. Diese werden oft als XML-Datei zum Download bereitgestellt. Viele der beschriebenen fiktionalen Welten haben ein größtenteils von Fans erstelltes Wiki. Durch einen XML-Dump lassen sich diese Wikis ohne Probleme mit der Software zur XML-Verarbeitung in Abschnitt 3.1.4 in Netze umwandeln. Durch ihr Erstellungsdatum sind die Dumps meist einfach eindeutig zu beschreiben.

Entdecken von Wandel

Durch die weite Verbreitung von Versionierung in Datenbeständen lassen sich ihre zeitlichen Zustände vergleichen. Die Suche über DBPedia im LWMap Service hat aufgezeigt, wie sich Cluster entdecken lassen. Hier könnte eine Zeitdimension ins Spiel gebracht werden. Die gleichen Anfragen auf dem gleichen Datensatz, zu verschiedenen Zeiten würden aufzeigen, wie sich die Dichte des Informationsnetzes in der Wikipedia zu bestimmten Themenbereichen verändert hat, kann ein Sachverhalt besser beleuchtet wurde und sich Kontexte zwischen den ausgewählten Knoten bildeten. Dies würde Aufschluss über die Konstruktion und die Entwicklung von Datenquellen wie Wikipedia im Kleinen geben. Beispielsweise könnten so Kampagnenerfolge in der Dateneingabe gemessen werden. Die-

[33] Alle anderen Produkte fallen unter die Bezeichnung "Legends", dabei werden Elemente aus diesem Korpus wiederverwendet.[248] Dadurch ergibt sich ein höherer Spielraum für die Ausgestaltung der Welt.

ses Verfahren wäre auch auf kleine Spezialwikis anwendbar.
Wandel kann auch in den Gesetzestexten untersucht werden, denn Gesetze sind nicht statisch, sondern werden verändert und sind insofern Versioniert. Dabei kann die Zunahme der Komplexität über die Zeit und die Versionen untersucht werden. Eine Auswahl an wirklich benötigten Urteilen und Referenzen in Fachtexten oder einfacher Recht zitierender Korrespondenz könnte einen Hinweis darauf geben, in welchen Bereichen Gesetzesänderungen zu einem Komplexitätsanstieg oder einer Reduktion geführt haben.

Literatur

[1] zuletzt aufgerufen 01.02.2016. Deutsches Archäologisches Institut. URL: https://gazetteer.dainst.org/.

[2] zuletzt aufgerufen 01.02.2016. Deutsches Archäologisches Institut. URL: http://zenon.dainst.org/.

[3] *A Song of Ice and Fire Wiki*. zuletzt aufgerufen 01.02.2016. URL: http://iceandfire.wikia.com/wiki/A_Song_of_Ice_and_Fire_Wiki.

[4] B. Adida und M. Birbeck. *RDFa Primer. Bridging the Human and Data Webs. W3C Working Group Note 14 October 2008*. 2008. URL: http://www.w3.org/TR/xhtml-rdfa-primer/.

[5] Guillaume Plique Alexis Jacomy. *Sigma js*. zuletzt aufgerufen 01.02.2016. URL: http://sigmajs.org/.

[6] José Ignacio Alvarez-Hamelin u. a. „K-core decomposition of Internet graphs: hierarchies, self-similarity and measurement biases.“ In: *NHM* 3.2 (2008), S. 371–393. URL: http://dblp.uni-trier.de/db/journals/nhm/nhm3.html#Alvarez-HamelinDBV08.

[7] AnonMoos. *Cretan-labyrinth-square*. Lizenz: Public domain zuletzt aufgerufen 01.02.2016. Wikimedia Commons. URL: https://commons.wikimedia.org/wiki/File:Cretan-labyrinth-square.svg.

[8] F. J. Anscombe. „Graphs in Statistical Analysis“. English. In: *The American Statistician* 27.1 (1973), pp. 17–21. ISSN: 00031305.

[9] *Apache Jena Fuseki*. zuletzt aufgerufen 01.02.2016. Apache Foundation. URL: https://jena.apache.org/documentation/fuseki2/index.html.

[10] *asm.js - an extraordinarily optimizable, low-level subset of JavaScript*. zuletzt aufgerufen 01.02.2016. URL: http://asmjs.org/.

[11] Sören Auer u. a. „DBpedia: A Nucleus for a Web of Open Data“. In: *In 6th Int'l Semantic Web Conference, Busan, Korea*. Springer, 2007, S. 11–15.

[12] Anatol Badach und Erwin Hoffmann. *Technik der IP-Netze. TCP/IP incl. IPv6 - Funktionsweise, Protokolle und Dienste*. 2., aktualisierte und erw. Aufl. Material: Elektronische Ressource ; Zugriff nur im Hochschulnetz der Universität Köln bzw. für autorisierte Benutzer ; Online-Ausg. ; ISBN der Parallelausgabe: 978-3-446-

21935-9 ; Quelldatenbank: KUGUSBJSON ; Format:marcform: print ; Umfang: XXIV, 696 S. : graph. Darst. München u.a.: Hanser, 2007. ISBN: 978-3-446-41089-3.

[13] L. Bagrow. *History of Cartography*. Transaction Publishers, 2009. ISBN: 9781412825184.

[14] Michael Balzer, Oliver Deussen und Claus Lewerentz. „Voronoi Treemaps for the Visualization of Software Metrics“. In: *Proceedings of the 2005 ACM Symposium on Software Visualization*. SoftVis '05. St. Louis, Missouri: ACM, 2005, S. 165–172. ISBN: 1-59593-073-6.

[15] A.-L. Barabási u. a. „Power-law distribution of the world wide web“. In: *Science* 287 (2000), 2115a.

[16] Albert-László Barabási und Réka Albert. „Emergence of scaling in random networks“. In: *science* 286.5439 (1999), S. 509–512.

[17] R.C. Barros, A.C.P.L.F. de Carvalho und A.A. Freitas. *Automatic Design of Decision-Tree Induction Algorithms*. SpringerBriefs in Computer Science. Springer International Publishing, 2015. ISBN: 9783319142319.

[18] Mathieu Bastian, Sebastien Heymann und Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. URL: `http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154`.

[19] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy u. a. „Gephi: an open source software for exploring and manipulating networks.“ In: *ICWSM* 8 (2009), S. 361–362.

[20] Vladimir Batagelj und Andrej Mrvar. „Some analyses of Erdös collaboration graph“. In: *Social Networks* 22.2 (2000), S. 173 –186. ISSN: 0378-8733.

[21] Shalini Batra und Charu Tyagi. „Comparative analysis of relational and graph databases“. In: *International Journal of Soft Computing and Engineering (IJSCE)* 2.2 (2012), S. 509–512.

[22] A.A. Bavelas. „A Mathematical Model for Group Structures“. In: *Human Organization* 7 (1948), S. 16–30.

[23] F. Bellomi und R. Bonato. „Network analysis for Wikipedia“. In: *Proceedings of Wikimania 2005, The First International Wikimedia Conference*.

[24] Sue Bennett, Karl Maton und Lisa Kervin. „The 'digital natives' debate: A critical review of the evidence“. In: *British journal of educational technology* 39.5 (2008), S. 775–786.

[25] T. Berners-Lee, J. Hendler und O. Lassila. „The semantic Web“. In: *Scientific American* 284.5 (2001), S. 28–37.

[26] Tim Berners-Lee. *Linked Data*. English. zuletzt aufgerufen 01.02.2016. Juli 2006. URL: `http://www.w3.org/DesignIssues/LinkedData.html`.

[27] Tim Berners-Lee und Dan Connolly. *Notation3 (N3): A readable RDF syntax*. Techn. Ber. W3C, Jan. 2008. URL: `http://www.w3.org/TeamSubmission/n3/`.

[28] Joanna Biega, Erdal Kuzey und Fabian M Suchanek. „Inside YAGO2s: A transparent information extraction architecture“. In: *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee. 2013, S. 325–328.

[29] Electronic Arts Binary Systems. *Starflight 2: Trade Routes of the Cloud Nebula*. Video Game. 1989.

[30] *Bing*. zuletzt aufgerufen 01.02.2016. Microsoft. URL: `http://www.bing.com/`.

[31] Christian Bizer. „D2r map-a database to rdf mapping language“. In: (2003).

[32] Christian Bizer und Richard Cyganiak. „D2r server-publishing relational databases on the semantic web“. In: *Poster at the 5th International Semantic Web Conference*. 2006.

[33] Christian Bizer u. a. „{DBpedia} - A crystallization point for the Web of Data“. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 7.3 (2009), S. 154 –165.

[34] V.D. Blondel u. a. „Fast unfolding of communities in large networks“. In: *J. Stat. Mech* (2008), P10008.

[35] Roberto Bonzio. *Father Busa, pioneer of computing in humanities with Index Thomisticus, dies at 98*. 2011. URL: `http://www.forbes.com/sites/robertobonzio/2011/08/11/father-busa-pioneer-of-computing-in-humanities-dies-at-98/` (besucht am 06. 01. 2016).

[36] S. Borgatti und M. Everett. „Network Analysis of 2-Mode Data“. In: *Social Networks* 19.3 (1997), S. 243–269.

[37] danah m. boyd und Nicole B. Ellison. „Social Network Sites: Definition, History, and Scholarship“. In: *Journal of Computer-Mediated Communication* 13.1 (2007), S. 210–230. ISSN: 1083-6101.

[38] Ulrik Brandes. „A faster algorithm for betweenness centrality*“. In: *The Journal of Mathematical Sociology* 25.2 (2001), S. 163–177. eprint: `http://www.tandfonline.com/doi/pdf/10.1080/0022250X.2001.9990249`.

[39] Ulrik Brandes, Markus Eiglsperger und Jürgen Lerner, Hrsg. *GraphML Primer*. Website. 2012. URL: `http://graphml.graphdrawing.org/primer/graphml-primer.html`.

[40] Ulrik Brandes und Thomas Erlebach. *Network analysis: methodological foundations*. Bd. 3418. Springer Science & Business Media, 2005.

[41] Ulrik Brandes und Dorothea Wagner. „A Bayesian paradigm for dynamic graph layout“. In: *Graph Drawing*. Springer. 1997, S. 236–247.

[42] Ronald L. Breiger. „The Duality of Persons and Groups“. In: *Social Forces* 53.2 (1974), S. 181–190. ISSN: 00377732.

[43] Ronald L Breiger und Philippa E Pattison. „Cumulated social roles: The duality of persons and their algebras“. In: *Social Networks* 8.3 (1986), S. 215 –256. ISSN: 0378-8733.

[44] Dan Brickley und Ramanathan Guha. *RDF Schema 1.1*. W3C Recommendation. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/. W3C, Feb. 2014.

[45] *Bundestagswahl - Bundestagswahl 2013 in der Stadt Köln - Wahlbezirke:* zuletzt aufgerufen 01.02.2016. Stadt Köln. 2013. URL: `http://www.stadt-koeln.de/wahlen/bundestagswahl/09-2013/Bundestagswahl_Uebersicht_stbz_Erststimmen.html`.

[46] *Bundestagswahl 2013 - Download Wahldaten*. zuletzt aufgerufen 01.02.2016. Landeshauptstadt Stuttgart, Statistisches Amt. URL: `http://www.stuttgart.de/item/show/504442`.

[47] Roberto Busa und Eduardo Bernot und Enrique Alarcón. *Index Thomisticus*. Online. URL: `http://www.corpusthomisticum.org/it/index.age` (besucht am 06. 01. 2016).

[48] Marco Büchler, Frederik Baumgardt und Benjamin Bock. „Von Platon zu Alexander dem Grossen - automatische Extraktion von Topic-Maps-basierten Assoziationsketten aus Sozialen Netzwerken der Antike“. In: *eDITion* (2010), S. 15–19.

[49] M.F. Calvi. *Antiquae Urbis Romae Cum Regionibus Simulachrum*. Froben, 1556. URL: `http://arachne.uni-koeln.de/books/FabioCalvo1532`.

[50] C.H. Cap. *Theoretische Grundlagen der Informatik*. Springer Vienna, 2013. ISBN: 9783709193297.

[51] Salvatore Catanese u. a. „Crawling Facebook for Social Network Analysis Purposes“. In: *CoRR* abs/1105.6307 (2011).

[52] Andrew H. Caudwell. „Gource: visualizing software version control history“. In: *SPLASH/OOPSLA Companion*. 2010, S. 73–74.

[53] Daniel Chang u. a. „Visualizing the Republic of Letters“. In: *Stanford: Stanford University. Retrieved April* 21 (2009), S. 2014.

[54] Hong woo Chun u. a. „Extraction of gene-disease relations from Medline using domain dictionaries and machine learning“. In: *Proc. PSB 2006*. 2006, S. 4–15.

[55] Reuven Cohen u. a. „Resilience of the Internet to Random Breakdowns“. In: *Phys. Rev. Lett.* 85 (21 2000), S. 4626–4628.

[56] J.S. Coleman und J.S. Coleman. *Foundations of Social Theory*. Belknap Series. Belknap Press of Harvard University Press, 1994. ISBN: 9780674312265.

[57] Christian Collberg u. a. „A system for graph-based visualization of the evolution of software“. In: *Proceedings of the 2003 ACM symposium on Software visualization*. ACM. 2003, 77–ff.

[58] *Comparison of SGML and XML*. zuletzt aufgerufen 01.02.2016. World Wide Web Consortium. 1997. URL: `https://www.w3.org/TR/NOTE-sgml-xml-971215/`.

[59] T.H. Cormen u. a. *Algorithmen - Eine Einführung*. Oldenbourg Wissenschaftsverlag, 2010. ISBN: 9783486590029.

[60] Kenneth A Cory. „Discovering hidden analogies in an online humanities database“. In: *Computers and the Humanities* 31.1 (1997), S. 1–12.

[61] Elizabeth Costenbader und Thomas W Valente. „The stability of centrality measures when networks are sampled“. In: *Social networks* 25.4 (2003), S. 283–307.

[62] Gregory Crane. „What do you do with a million books?“ In: *D-Lib magazine* 12.3 (2006), S. 1.

[63] Douglas Crockford. *Introducing JSON*. zuletzt aufgerufen 01.02.2016. 2002. URL: `https://commons.wikimedia.org/wiki/File\%3AMaze_simple.svg` (besucht am 24. 03. 2014).

[64] S. Cuy und L. Klafki. „Römer im Computer. Der Einsatz von GOLEM II im Forschungsarchiv für Antike Plasik 1974“. In: Hrsg. von P. Scheding und M. Remmy. Bd. Antike Plastik 5.0://: 50 Jahre Forschungsarchiv für Antike Plastik in Köln. Ausstellung im Akademischen Kunstmuseum, Antikensammlung der Universität Bonn, 26.10.2014 - 21.12.2014. Lit Verlag, 2014. Kap. 6, S. 68–75. ISBN: 9783643128126.

[65] Sebastian Cuy. „Automatische Metadatenextraktion aus archäologischen Fachpublikationen“. Magisterarb. Universität zu Köln, 2010.

[66] Luca Dall'Asta u. a. „Exploring networks with traceroute-like probes: Theory and simulations“. In: *Theoretical Computer Science* 355.1 (2006), S. 6–24.

[67] Jeffrey Dean und Sanjay Ghemawat. „MapReduce: simplified data processing on large clusters“. In: *Communications of the ACM* 51.1 (2008), S. 107–113.

[68] Franz-Benno Delonge. *Trans-America*. Brettspiel. Verlag: Winning Moves. 2002.

[69] Erwan Demairy. *Gephi - SemanticWebImport*. zuletzt aufgerufen 01.02.2016. 2013. URL: `https://marketplace.gephi.org/plugin/semanticwebimport/`.

[70] DMA Design. *Grand Theft Auto*. Video Game. 1997. URL: `http://www.rockstargames.com/gta/`.

[71] Stephan Diehl und Carsten Görg. „Graphs, they are changing“. In: *Graph drawing*. Springer. 2002, S. 23–31.

[72] Jana Diesner und Kathleen M Carley. „Revealing social structure from texts: metamatrix text analysis as a novel method for network text analysis“. In: *Causal mapping for information systems and technology research: Approaches, advances, and illustrations* (2005), S. 81–108.

[73] Jana Diesner und KathleenM. Carley. „Relationale Methoden in der Erforschung, Ermittlung und Prävention von Kriminalität“. German. In: *Handbuch Netzwerkforschung*. Hrsg. von Christian Stegbauer und Roger Häußling. VS Verlag für Sozialwissenschaften, 2010, S. 725–738. ISBN: 978-3-531-15808-2.

[74] Jana Diesner u. a. *Using Network Text Analysis to Detect the Organizational Structure of Covert Networks," presented at NAACSOS*. 2004.

[75] C. Doctorow. *Down And Out In The Magic Kingdom:* CreateSpace Independent Publishing Platform, 2014. ISBN: 9781497352292.

[76] Martin Doerr. *CIDOC Conceptual Reference Model (CRM)*. ISO. CIDOC Documentation Standards Working Group. URL: `http://www.cidoc-crm.org/`.

[77] Michael Seiferle Dr. Christian Grün Dr. Alexander Holupirek. *BaseX — The XML Database*. zuletzt aufgerufen 01.02.2016. URL: `http://basex.org/`.

[78] *Drupal - Open Source CMS*. zuletzt aufgerufen 01.02.2016. URL: `https://www.drupal.org/`.

[79] M. Duerst und M. Suignard. *RFC 3987: Internationalized Resource Identifiers (IRIs)*. RFC 3987 (Proposed Standard), see `http://www.ietf.org/rfc/rfc3987.txt`. Internet Engineering Task Force, 2005. URL: `http://www.ietf.org/rfc/rfc3987.txt`.

[80] Peter Dömel. „WebMap: a graphical hypertext navigation tool“. In: *Computer Networks and {ISDN} Systems* 28.1–2 (1995). Selected Papers from the Second World-Wide Web Conference, S. 85 –97. ISSN: 0169-7552.

[81] Marten Düring. *From Hermeneutics to Data to Networks: Data Extraction and Network Visualization of Historical Sources*. zuletzt aufgerufen 20.03.2015. 2015. URL: `http://programminghistorian.org/lessons/creating-network-diagrams-from-historical-sources`.

[82] Silke Eckstein. „Informationsmanagement in der Systembiologie. Datenbanken, Integration, Modellierung“. In: Springer, 2011. Kap. 6.

[83] O. A. W. Dilke with additional material supplied by the editors. „Vol. 1: Cartography in prehistoric, ancient, and medieval Europe and the Mediterranean“. In: Hrsg. von J Harley und D Woodward. Chicago: University of Chicago Press, 1987. Kap. Chapter 11: The Culmination of Greek Cartography in Ptolemy, S. 177–200.

[84] T.M. Egyedi und A.G.A.J. Loeffen. „Succession in standardization: grafting {XML} onto {SGML}“. In: *Computer Standards & Interfaces* 24.4 (2002). {XML} diffusion: transfer and differentiation, S. 279 –290. ISSN: 0920-5489.

[85] K. Elangovan. *GIS: Fundamentals, Applications and Implementations*. New India Publishing Agency, 2006. ISBN: 9788189422165.

[86] David K Elson, Nicholas Dames und Kathleen R McKeown. „Extracting social networks from literary fiction“. In: *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics. 2010, S. 138–147.

[87] Sierra Entertainment. *Quest for Glory II: Trial by Fire*. Video Game. 1990.

[88] Güler Ergün. „Human sexual contact network as a bipartite graph“. In: *Physica A: Statistical Mechanics and its Applications* 308.1–4 (2002), S. 483 –488. ISSN: 0378-4371.

[89] Martin Ester u. a. „A density-based algorithm for discovering clusters in large spatial databases with noise“. In: AAAI Press, 1996, S. 226–231.

[90] Stephen Eubank u. a. „Modelling disease outbreaks in realistic urban social networks“. In: *Nature* 429.6988 (2004), S. 180–184.

[91] *Facebook Quarterly Earnings Slides Q3 2013*. zuletzt aufgerufen 01.02.2016. Facebook. 2013. URL: http://allfacebook.de/wp-content/uploads/2013/11/FB_Q313InvestorDeck.pdf.

[92] Fanblade. *Graph of connections between territories on the Risk game board*. URL: https://upload.wikimedia.org/wikipedia/commons/2/24/Risk_Game_Graph.svg.

[93] Usama Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth. „From data mining to knowledge discovery in databases“. In: *AI magazine* 17.3 (1996), S. 37.

[94] Santo Fortunato. „Community detection in graphs“. In: *Physics Reports* 486.3–5 (2010), S. 75 –174. ISSN: 0370-1573.

[95] Allen Foster und Nigel Ford. „Serendipity and information seeking: an empirical study“. In: *Journal of Documentation* 59.3 (2003), S. 321–340.

[96] L.C. Freeman. „Centrality in Social Networks: Conceptual Clarification“. In: *Social Networks* 1 (1979), S. 215–239.

[97] Linton C Freeman. „Uncovering organizational hierarchies“. In: *Computational & Mathematical Organization Theory* 3.1 (1997), S. 5–18.

[98] ThomasN. Friemel und Andrea Knecht. „Praktikable vs. tatsächliche Grenzen von sozialen Netzwerken. Eine Diskussion zur Validität von Schulklassen als komplette Netzwerke“. German. In: *Grenzen von Netzwerken*. Hrsg. von Roger Häußling. VS Verlag für Sozialwissenschaften, 2009, S. 15–32. ISBN: 978-3-531-16308-6.

[99] Thomas M. J. Fruchterman, Edward und Edward M. Reingold. *Graph Drawing by Force-directed Placement*. 1991.

[100] Reinhard Förtsch. *Objektdatenbank und kulturelle Archive des Archäologischen Instituts der Universität zu Köln und des Deutschen Archäologischen Instituts*. zuletzt aufgerufen 01.02.2016. Forschungsarchiv für Antike Plastik. URL: http://www.arachne.uni-koeln.de/drupal/?q=de/node/3.

[101] Reinhard Förtsch und Rasmus Krempel. *Improving the Arachne-Pleiades matching*. en. zuletzt aufgerufen 01.02.2017. Juni 2012. URL: http://commons.pelagios.org/2012/06/improving-the-arachne-pleiades-matching/.

[102] *Game of Thrones Wiki*. zuletzt aufgerufen 01.02.2016. URL: http://gameofthrones.wikia.com/wiki/Game_of_Thrones_Wiki.

[103] *Gesetze im Internet*. zuletzt aufgerufen 01.02.2016. Bundesministerium der Justiz und für Verbraucherschutz. URL: http://www.gesetze-im-internet.de/.

[104] *GEXF 1.2draft Primer*. zuletzt aufgerufen 01.02.2016. GEXF Working Group. 2012. URL: https://gephi.org/gexf/1.2draft/gexf-12draft-primer.pdf.

[105] James J Gibson. „The ecological approach to the visual perception of pictures“. In: *Leonardo* 11.3 (1978), S. 227–235.

[106] *Global Language Network*. zuletzt aufgerufen 18.09.2015. URL: http://language.media.mit.edu/visualizations/wikipedia.

[107] Erving Goffman u. a. „The presentation of self in everyday life“. In: (1959).

[108] E Bruce Goldstein. „The ecology of JJ Gibson's perception“. In: *Leonardo* 14.3 (1981), S. 191–195.

[109] *Google Maps*. zuletzt aufgerufen 01.02.2016. Google. URL: http://maps.google.com/.

[110] *Google Ngram Viewer*. zuletzt aufgerufen 01.02.2016. Google. URL: https://books.google.com/ngrams.

[111] Brendan Griffen. *Graphs of Wikipedia: Influential Thinkers*. 2013. URL: http://brendangriffen.com/gow-influential-thinkers/.

[112] W3C RDF Data Access Working Group. *SPARQL Query Language for RDF*. zuletzt aufgerufen 01.02.2016. W3C. 2008. URL: https://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/.

[113] Christian Grün. „Pushing XML Main Memory Databases to their Limits.“ In: *Grundlagen von Datenbanken*. 2006, S. 60–64.

[114] Jiancheng Guan und Zifeng Chen. „Patent collaboration and international knowledge flow“. In: *Information Processing & Management* 48.1 (2012), S. 170 –181. ISSN: 0306-4573.

[115] Frank Gurski u. a. *Exakte Algorithmen für schwere Graphenprobleme*. eXamen.press. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2010.

[116] Kevin Guthrie, Rebecca Griffiths und Nancy Maron. „Sustainability and revenue models for online academic resources“. In: *An Ithaka Report* (2008).

[117] R.H. Güting und S. Dieker. *Datenstrukturen und Algorithmen*. XLeitfäden der Informatik. Vieweg+Teubner Verlag, 2013. ISBN: 9783322891136.

[118] Rasmus Hahn u. a. „Faceted Wikipedia Search“. English. In: *Business Information Systems*. Hrsg. von Witold Abramowicz und Robert Tolksdorf. Bd. 47. Lecture Notes in Business Information Processing. Springer Berlin Heidelberg, 2010, S. 1–11. ISBN: 978-3-642-12813-4.

[119] M. Haklay und P. Weber. „OpenStreetMap: User-Generated Street Maps“. In: *Pervasive Computing, IEEE* 7.4 (2008), S. 12–18. ISSN: 1536-1268.

[120] Borgatti Halgin. „An Introduction to Personal Network Analysis and Tie Churn Statistics Using E-NET“. In: *Connections* 32.1 (2012), S. 37–49.

[121] Garrett Hardin. „The tragedy of the commons“. In: *science* 162.3859 (1968), S. 1243–1248.

[122] P. D. A. Harvey. „Vol. 1: Cartography in prehistoric, ancient, and medieval Europe and the Mediterranean“. In: Hrsg. von J Harley und D Woodward. Chicago: University of Chicago Press, 1987. Kap. Chapter 20: Local and Regional Cartography in Medieval Europe, S. 464–501.

[123] Brian Hayes u. a. „First links in the Markov chain“. In: *American Scientist* 101.2 (2013), S. 92.

[124] DanaL. Haynie. „Friendship Networks and Delinquency: The Relative Nature of Peer Delinquency“. English. In: *Journal of Quantitative Criminology* 18.2 (2002), S. 99–134. ISSN: 0748-4518.

[125] Hazard-SJ. *Wikidata:Bots*. zuletzt aufgerufen 01.02.2016. Wikidata. 2015. URL: `https://www.wikidata.org/w/index.php?title=Wikidata:Bots&oldid=285391005`.

[126] X. He und S. Xu. *Process Neural Networks: Theory and Applications*. Advanced Topics in Science and Technology in China. Springer Berlin Heidelberg, 2010. ISBN: 9783540737629.

[127] Marti Hearst. „Design recommendations for hierarchical faceted search interfaces“. In: *ACM SIGIR workshop on faceted search*. Seattle, WA. 2006, S. 1–5.

[128] Tom Heath und Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space*. 1. Aufl. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool, 2011. ISBN: 9781608454303 (paperback), 9781608454310 (ebook).

[129] Fritz Heider. „Attitudes and Cognitive Organization“. In: *The Journal of Psychology* 21.1 (1946). PMID: 21010780, S. 107–112. eprint: `http://www.tandfonline.com/doi/pdf/10.1080/00223980.1946.9917275`.

[130] J. E. Hirsch. „An index to quantify an individual's scientific research output“. In: *Proceedings of the National Academy of Sciences of the United States of America* 102.46 (2005). eprint: `http://www.pnas.org/content/102/46/16569.full.pdf+html`.

[131] Bernie Hogan. zuletzt aufgerufen 18.09.2015. URL: `http://wikiproject.oii.ox.ac.uk/networks/index.php`.

[132] Yifan Hu. „Efficient, high-quality force-directed graph drawing“. In: *Mathematica Journal* 10.1 (2005), S. 37–71.

[133] Darrell Huff. „Howto Lie With Statistics“. In: *New York1954* (1954).

[134] Ullrich Hustadt. *Do We Need the Closed-World Assumption in Knowledge Representation?* 1994.

[135] *IMDb - Movies, TV and Celebrities*. URL: `http://www.imdb.com/` (besucht am 06. 01. 2016).

[136] Leif Isaksen. „The application of network analysis to ancient transport geography: A case study of Roman Baetica“. In: *Digital Medievalist* 4 (2008).

[137] Leif Isaksen u. a. „Pelagios and the emerging graph of ancient world data“. In: *Proceedings of the 2014 ACM conference on Web science*. ACM. 2014, S. 197–201.

[138] Jkwchui. *Maze simple*. Lizenz CC0 zuletzt aufgerufen 01.02.2016. URL: `https://commons.wikimedia.org/wiki/File\%3AMaze_simple.svg`.

[139] David Easley Jon Kleinberg. *Networks, Crowds, and Markets ; Reasoning about a Highly Connected World*. [S.l.]: Cambridge University Press, 2010. ISBN: 9780521195331 0521195330.

[140] Denise B. Kandel. „Homophily, Selection, and Socialization in Adolescent Friendships“. English. In: *American Journal of Sociology* 84.2 (1978), pp. 427–436. ISSN: 00029602.

[141] Connie Kasari u. a. „Social Networks and Friendships at School: Comparing Children With and Without ASD“. English. In: *Journal of Autism and Developmental Disorders* 41.5 (2011), S. 533–544. ISSN: 0162-3257.

[142] Michael Kifer. „Rule Interchange Format: The Framework“. English. In: *Web Reasoning and Rule Systems*. Hrsg. von Diego Calvanese und Georg Lausen. Bd. 5341. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, S. 1–11. ISBN: 978-3-540-88736-2.

[143] Rebecca Sutton Koeser. *Using DBpedia to graph writers influence*. 2012. URL: `http://disc.library.emory.edu/networkingbelfast/dbpedia-writers-influence/`.

[144] Michal Konkol, Tomáš Brychcín und Miloslav Konopík. „Latent semantics in Named Entity Recognition“. In: *Expert Systems with Applications* 42.7 (2015), S. 3470 –3479. ISSN: 0957-4174.

[145] Gueorgi Kossinets. „Effects of missing data in social networks“. In: *Social networks* 28.3 (2006), S. 247–268.

[146] Rasmus Krempel. *An Update for the Arachne-Pleiades annotations*. en. zuletzt aufgerufen 01.02.2017. 2011. URL: `http://commons.pelagios.org/2011/08/an-update-for-the-arachne-pleiades-annotations/`.

[147] Rasmus Krempel. *Creating the Pleiades to Arachne annotation*. en. zuletzt aufgerufen 01.02.2017. 2011. URL: `http://commons.pelagios.org/2011/07/creating-the-pleiades-to-arachne-annotation/`.

[148] Markus Krötzsch, Denny Vrandečić und Max Völkel. „Semantic MediaWiki“. English. In: *The Semantic Web - ISWC 2006*. Hrsg. von Isabel Cruz u. a. Bd. 4273. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2006, S. 935–942. ISBN: 978-3-540-49029-6.

[149] Hartmut Kugler. *Die Ebstorfer Weltkarte*. Website. URL: `http://www.leuphana.de/ebskart`.

[150] Bill Kules u. a. „What Do Exploratory Searchers Look at in a Faceted Search Interface?“ In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '09. Austin, TX, USA: ACM, 2009, S. 313–322. ISBN: 978-1-60558-322-8.

[151] Robert Kummer. *Intelligent information access to linked data. weaving the cultural heritage web*. 2013.

[152] Albert Lamorisse. *Risiko*. Schmidt Spiele. 1961.

[153] David Laniado u. a. „When the wikipedians talk: Network and tree structure of wikipedia discussion pages.“ In: *ICWSM*. 2011.

[154] J. Latimer. *The Lady in the Morgue*. A Bill Crane Mystery. Orion, 2013. ISBN: 9781471910715.

[155] Ramon Lawrence. „The space efficiency of {XML}“. In: *Information and Software Technology* 46.11 (2004), S. 753 –759. ISSN: 0950-5849.

[156] Berners T. Lee, L. Masinter und M. Mccahill. *RFC 1738: Uniform Resource Locator (URL)*. http://www.ietf.org/rfc/rfc1738.txt. 1994. URL: `http://www.ietf.org/\-rfc/\-rfc1738.txt`.

[157] Fredrik Liljeros, Christofer R. Edling und Luis A.Nunes Amaral. „Sexual networks: implications for the transmission of sexually transmitted infections“. In: *Microbes and Infection* 5.2 (2003), S. 189 –196. ISSN: 1286-4579.

[158] H.P. Lovecraft. „The Call of Cthulhu“. In: *Weird Tales* 11.2 (1928).

[159] H.P. Lovecraft. „The Dreams in the Witch House“. In: *Weird Tales* (1933). URL: `https://en.wikisource.org/wiki/The_Dreams_in_the_Witch-House`.

[160] *Madonna swears at music pirates*. zuletzt aufgerufen 01.02.2016. BBC. 2003. URL: `http://news.bbc.co.uk/2/hi/2962475.stm`.

[161] Shawn Martin u. a. „OpenOrd: an open-source toolbox for large graph layout“. In: *IS&T/SPIE Electronic Imaging*. International Society for Optics und Photonics. 2011.

[162] Sorin Matei. „Analyzing Social Media Networks with NodeXL: Insights from a Connected World“. In: *Intl. Journal of Human–Computer Interaction* 27.4 (2011), S. 405–408.

[163] WarrenS. McCulloch und Walter Pitts. „A logical calculus of the ideas immanent in nervous activity“. English. In: *The bulletin of mathematical biophysics* 5.4 (1943), S. 115–133. ISSN: 0007-4985.

[164] Jean M. McGloin. *Street Gangs and Interventions: Innovative Problem Solving with Network Analysis*. US Department of Justice, 2005.

[165] Deborah L. McGuinness und Frank van Harmelen. *OWL Web Ontology Language Overview*. W3C Recommendation. World Wide Web Consortium, 2004. URL: `http://www.w3.org/TR/owl-features`.

[166] Wolfgang et al. Meier. *eXistdb - The Open Source Native XML Database*. zuletzt aufgerufen 01.02.2016. URL: `http://exist-db.org/`.

[167] Pablo N. Mendes u. a. „DBpedia spotlight: shedding light on the web of documents“. In: *Proceedings of the 7th International Conference on Semantic Systems*. I-Semantics '11. Graz, Austria: ACM, 2011, S. 1–8. ISBN: 978-1-4503-0621-8.

[168] Robert K. Merton. „The Role-Set: Problems in Sociological Theory“. In: *The British Journal of Sociology* 8.2 (1957), S. 106–120. ISSN: 00071315, 14684446.

[169] Peter Mika. „Ontologies are us: A unified model of social networks and semantics“. In: *Web Semantics: Science, Services and Agents on the World Wide Web* 5.1 (2007). Selected Papers from the International Semantic Web Conference {ISWC} 2005 International Semantic Web Conference (ISWC2005), S. 5 –15. ISSN: 1570-8268.

[170] Stanley Milgram. „The Small World Problem“. In: *Psychology Today* 1.1 (1967), S. 61–67.

[171] George Miller. *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*. 1956. URL: `http://cogprints.org/730/`.

[172] Jack Minker. „On indefinite databases and the closed world assumption“. English. In: *6th Conference on Automated Deduction*. Hrsg. von D.W. Loveland. Bd. 138. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 1982, S. 292–308. ISBN: 978-3-540-11558-8.

[173] Mark Monmonier. *How to lie with maps*. Chicago: University of Chicago Press, 1996. ISBN: 978-0-226-53421-3.

[174] Gordon E. Moore. „Cramming more components onto integrated circuits“. In: *Electronics* 38.8 (1965).

[175] *Multi-Agent Transport Simulation — MATSim*. zuletzt aufgerufen 01.02.2016. URL: `http://matsim.org/`.

[176] Paul Mutton. *PieSpy Social Network Bot*. zuletzt aufgerufen 01.02.2016. URL: `http://www.jibble.org/piespy/`.

[177] Paul Mutton. *Shakespeare Social Networks*. zuletzt aufgerufen 01.02.2016. URL: `http://www.jibble.org/shakespeare/`.

[178] David Nadeau und Satoshi Sekine. „A survey of named entity recognition and classification“. In: *Lingvisticae Investigationes* 30.1 (2007-01-01T00:00:00), S. 3–26.

[179] HB Newcombe u. a. „Automatic linkage of vital records“. In: *Record linkage techniques, 1985: proceedings of the Workshop on Exact Matching Methodologies, Arlington, Virginia, May 9-10, 1985: co-sponsored with the Washington Statistical Society and the Federal Committee on Statistical Methodology*. Bd. 1299. Dept. of the Treasury, Internal Revenue Service, Statistics of Income Division. 1986, S. 7.

[180] M. E. J. Newman. „Analysis of weighted networks“. In: *Phys. Rev. E* 70 (5 2004), S. 056131.

[181] M.E.J. Newman. „Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality“. In: *Phys. Rev. E* 64.1 (Juni 2001).

[182] *Nike von Samothrake*. zuletzt aufgerufen 01.02.2016. Arachne Datenbank. URL: `http://arachne.uni-koeln.de/item/reproduktion/3000339`.

[183] *Nike von Samothrake*. zuletzt aufgerufen 01.02.2016. Arachne Datenbank. URL: `http://arachne.uni-koeln.de/item/reproduktion/3316013`.

[184] Manfred Nitzsche. *Graphen für Einsteiger. Rund um das Haus vom Nikolaus*. 3., überarbeitete und erweiterte Auflage. Material: Elektronische Ressource ; Zugriff nur im Hochschulnetz der Universität Köln bzw. für autorisierte Benutzer ; ISBN der Parallelausgabe: 978-3-8348-0813-4 ; Quelldatenbank: KUGUSBJSON. Wiesbaden: Vieweg+Teubner Verlag / GWV Fachverlage GmbH, Wiesbaden, 2009. ISBN: 978-3-8348-9968-2.

[185] Arlind Nocaj, Mark Ortmann und Ulrik Brandes. „Untangling Hairballs“. In: *Graph Drawing*. Springer. 2014, S. 101–112.

[186] *Node.js*. zuletzt aufgerufen 01.02.2016. URL: `https://nodejs.org/`.

[187] Jong-Hoon Oh u. a. „Enriching multilingual language resources by discovering missing cross-language links in Wikipedia“. In: *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society. 2008, S. 322–328.

[188] *Open Annotation Collaboration*. zuletzt aufgerufen 01.02.2016. W3C Open Annotation Community Group. URL: `http://www.openannotation.org/`.

[189] *Open Street Maps*. zuletzt aufgerufen 01.02.2016. URL: `http://www.openstreetmap.org/`.

[190] *OpenLayers 3 - A high-performance, feature-packed library for all your mapping needs.* zuletzt aufgerufen 26.02.2015. URL: `http://www.openlayers.org/`.

[191] LordT Orthuberra Fluteflute. *Risk map in Wikipedia*. 2009. URL: `https://upload.wikimedia.org/wikipedia/commons/9/9d/Risk_game_map.png`.

[192] John F. Padgett und Christopher K. Ansell. „Robust Action and the Rise of the Medici, 1400-1434“. English. In: *American Journal of Sociology* 98.6 (1993), pp. 1259–1319. ISSN: 00029602.

[193] Lawrence Page u. a. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford InfoLab, 1999.

[194] Anuška Ferligoj Patrick Doreian Vladimir Batagelj. *Pajek data: Generalized Blockmodeling / Davis*. zuletzt aufgerufen 01.02.2016. 2007. URL: `http://vlado.fmf.uni-lj.si/pub/networks/data/GBM/davis.htm`.

[195] Denise Peikert. *Das ist unser Rockerkrieg*. zuletzt aufgerufen 01.02.2016. FAZ. 2013. URL: http://www.faz.net/aktuell/rhein-main/hells-angels-verbot-das-ist-unser-rockerkrieg-12156306.html.

[196] S. Pemberton u. a. *XHTML 1.0 The Extensible HyperText Markup Language*. Hrsg. von S. Pemberton et al. Second. W3C Recommendation. 2002. URL: http://www.w3.org/TR/xhtml1.

[197] Jean Piaget. *Die Entwicklung der elementaren logischen Strukturen*. 1. Aufl. Sprache und Lernen, Teil 32. Quelldatenbank: KUGJSON ; Format:marcform: print ; Umfang: 209 S. : graph. Darst. 1973. ISBN: 3-7895-0235-9.

[198] A. Platzer und S. Herz. *Application programming interfaces for scrolling operations*. US Patent 7,844,915. 2010. URL: https://www.google.de/patents/US7844915.

[199] *Pleiades*. zuletzt aufgerufen 01.02.2016. neh.gov — National Endowment for the Humanities. URL: http://pleiades.stoa.org/.

[200] JA Pouwelse u. a. *A measurement study of the bittorrent peer-to-peer file-sharing system*. Techn. Ber. Technical Report PDS-2004-003, Delft University of Technology, The Netherlands, 2004.

[201] David Martin Powers. „Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation“. In: (2011).

[202] Marc Prensky. „Digital natives, digital immigrants part 1“. In: *On the horizon* 9.5 (2001), S. 1–6.

[203] *Preventing comment spam*. 00005. Jan. 2005. URL: https://googleblog.blogspot.com/2005/01/preventing-comment-spam.html (besucht am 06. 01. 2016).

[204] *Program Package Pajek / PajekXXL*. zuletzt aufgerufen 01.02.2016. URL: http://pajek.imfm.si/doku.php.

[205] *QGIS Ein freies Open-Source-Geographisches-Informationssystem*. URL: http://www.qgis.org/ (besucht am 06. 01. 2016).

[206] Cartic Ramakrishnan, Krys J Kochut und Amit P Sheth. „A framework for schema-driven relationship discovery from unstructured text“. In: *The Semantic Web-ISWC 2006*. Springer, 2006, S. 583–596.

[207] Simon Raper. *Graphing the history of philosophy*. Blog. 2012. URL: http://drunks-and-lampposts.com/2012/06/13/graphing-the-history-of-philosophy/.

[208] A. Recht und F. Schäfer. „Tradition und Innovation 3.0:// Das Forschungsarchiv in den Jahren 1990-2000“. In: Hrsg. von P. Scheding und M. Remmy. Bd. Antike Plastik 5.0://: 50 Jahre Forschungsarchiv für Antike Plastik in Köln. Ausstellung im Akademischen Kunstmuseum, Antikensammlung der Universität Bonn, 26.10.2014 - 21.12.2014. Lit Verlag, 2014. Kap. 3, S. 33–42. ISBN: 9783643128126.

[209] A Renyi und P Erdos. „On random graphs“. In: *Publicationes Mathematicae* 6.290-297 (1959), S. 5.

[210] M. Rieser und K. Nagel. „Network breakdown “at the edge of chaos” in multi-agent traffic simulations“. English. In: *The European Physical Journal B* 63.3 (2008), S. 321–327. ISSN: 1434-6028.

[211] Colin Ritchie. *Relational Database Principles*. Cengage Learning EMEA, 2002.

[212] I. Robinson, J. Webber und E. Eifrem. *Graph Databases*. O'Reilly Media, Incorporated, 2013. ISBN: 9781449356262.

[213] Shahar Ronen u. a. „Links that speak: The global language network and its association with global fame“. In: *Proceedings of the National Academy of Sciences* 111.52 (2014), E5616–E5622.

[214] Jeff Rydberg-Cox. „Social networks and the language of greek tragedy“. In: *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*. Bd. 1. 3. 2011.

[215] Gerard Salton und Christopher Buckley. „Term-weighting approaches in automatic text retrieval“. In: *Information processing & management* 24.5 (1988), S. 513–523.

[216] Sebastian Schaffert u. a. „Semantic Wiki“. German. In: *Informatik-Spektrum* 30.6 (2007), S. 434–439. ISSN: 0170-6012.

[217] P. Scheding und M. Remmy. *Antike Plastik 5.0://: 50 Jahre Forschungsarchiv für Antike Plastik in Köln. Ausstellung im Akademischen Kunstmuseum, Antikensammlung der Universität Bonn, 26.10.2014 - 21.12.2014*. Lit Verlag, 2014. ISBN: 9783643128126.

[218] W. Scheidel und E. Meeks. *ORBIS: The Stanford Geospatial Network Model of the Roman World*. zuletzt aufgerufen 01.02.2016. 2012. URL: `http://orbis.stanford.edu`.

[219] M. Schich. *Rezeption und Tradierung als komplexes Netzwerk*. Biering & Brinkmann, 2009. ISBN: 9783930609567.

[220] Maximilian Schich u. a. „A network framework of cultural history“. In: *science* 345.6196 (2014), S. 558–562.

[221] Sebastian Cuy – Philipp Gerth – Maximilian Heiden – Wibke Kolbmann – Wolfgang Schmidle. „iDAI.gazetteer – ein Referenzsystem für altertumswissenschaftliche Ortsinformationen als Teil einer digitalen Forschungsinfrastruktur“. In: *In Kölner und Bonner Archaeologica* (2014).

[222] Raoul Schönhof, Axel Tenschert und Alexey Cheptsov. „Towards Legal Knowledge Representation System Leveraging RDF“. In: *SEMAPRO 2014, The Eighth International Conference on Advances in Semantic Processing*. 2014, S. 13–16.

[223] Karen Schwane. *Will in Town Ein Location-based Game, entwickelt von Studierenden des Instituts für Medienkultur und Theater*. zuletzt aufgerufen 01.02.2016. Fachschaft Medienwissenschaften an der Universität zu Köln c/o Institut für Medienkultur und Theater. URL: `http://www.medienwissenschaften.net/magazin/alle-artikel/will-in-town/`.

[224] Nigel Shadbolt, Tim Berners-Lee und Wendy Hall. „The Semantic Web Revisited“. In: *IEEE Intelligent Systems* 21.3 (2006), S. 96–101.

[225] *Shakespeare's Works in TEI*. zuletzt aufgerufen 01.02.2016. eXistdb. URL: http://exist-db.org/exist/apps/shakespeare/works/.

[226] Paul Shannon u. a. „Cytoscape: a software environment for integrated models of biomolecular interaction networks“. In: *Genome research* 13.11 (2003), S. 2498–2504.

[227] *Shape Stimmbezirk*. zuletzt aufgerufen 01.02.2016. Offene Daten Köln. 2013. URL: http://www.offenedaten-koeln.de/dataset/7ea6c349-84bc-4a15-8c58-d4cd4d5de1f5/resource/7ea6c349-84bc-4a15-8c58-d4cd4d5de1f5.

[228] Jianhong Shen. „On the foundations of vision modeling: I. Weber's law and Weberized {TV} restoration“. In: *Physica D: Nonlinear Phenomena* 175.3–4 (2003), S. 241 –251. ISSN: 0167-2789.

[229] Ben Shneiderman. „Tree Visualization with Tree-maps: 2-d Space-filling Approach“. In: *ACM Trans. Graph.* 11.1 (Jan. 1992), S. 92–99. ISSN: 0730-0301.

[230] Herbert A Simon. „Designing organizations for an information-rich world“. In: *Computers, communication, and the public interest* 37 (1971), S. 40–41.

[231] Rainer Simon, Elton Barker und Leif Isaksen. „Exploring Pelagios: a visual browser for geo-tagged datasets“. In: (2012).

[232] S. Singh. *The Simpsons and Their Mathematical Secrets*. Bloomsbury Publishing Plc, 2013. ISBN: 9781408835302.

[233] Evren Sirin und Bijan Parsia. „SPARQL-DL: SPARQL Query for OWL-DL.“ In: *OWLED*. Bd. 258. 2007.

[234] Andre Skupin. „From Metaphor to Method: Cartographic Perspectives on Information Visualization“. In: *Proceedings of InfoVis 2000 (Salt Lake City UT)* (2000), S. 91–98.

[235] *Software Library: MS-DOS Games*. zuletzt aufgerufen 01.02.2016. Internet Archive. URL: https://archive.org/details/softwarelibrary_msdos_games&tab=about.

[236] Daniel AJ Sokolov. „MP3-Player: Microsoft macht dem Zune den Garaus“. In: *Heise Online* (16.11.2015). URL: http://www.heise.de/newsticker/meldung/MP3-Player-Microsoft-macht-dem-Zune-den-Garaus-2921846.html.

[237] Philipp Sorg und Philipp Cimiano. „Enriching the crosslingual link structure of wikipedia-a classification-based approach“. In: *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*. 2008, S. 49–54.

[238] Andreas Stanescu. „Assessing the durability of formats in a digital preservation environment“. In: *OCLC Systems & Services: International digital library perspectives* 21.1 (2005), S. 61–81. eprint: http://dx.doi.org/10.1108/10650750510578163.

[239] Pontus Stenetorp u. a. „BRAT: A Web-based Tool for NLP-assisted Text Annotation“. In: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. EACL '12. Avignon, France: Association for Computational Linguistics, 2012, S. 102–107.

[240] James Stiller, Daniel Nettle und Robin IM Dunbar. „The small world of Shakespeare's plays“. In: *Human Nature* 14.4 (2003), S. 397–408.

[241] *Stop, reader! Fr Busa is dead*. En. Zeitung. Aug. 2011. URL: http://www.osservatoreromano.va/en/news/stop-reader-fr-busa-is-dead (besucht am 06. 01. 2016).

[242] Florian Straus. „Netzwerkkarten – Netzwerke sichtbar machen“. German. In: *Handbuch Netzwerkforschung*. Hrsg. von Christian Stegbauer und Roger Häußling. VS Verlag für Sozialwissenschaften, 2010, S. 527–538. ISBN: 978-3-531-15808-2.

[243] Roberto Tamassia. *Handbook of graph drawing and visualization*. CRC press, 2013.

[244] Ling-Xiang Tang u. a. „An evaluation framework for cross-lingual link discovery“. In: *Information Processing & Management* 50.1 (2014), S. 1–23.

[245] NWB Team. *Network Workbench Tool*. zuletzt aufgerufen 01.02.2016. Indiana University, Northeastern University, und University of Michigan. 2006. URL: http://nwb.slis.indiana.edu.

[246] Sci2 Team. *Science of Science (Sci2) Tool*. zuletzt aufgerufen 01.02.2016. Indiana University und SciTech Strategies. 2009. URL: https://sci2.cns.iu.edu.

[247] *TEI: Text Encoding Initiative*. zuletzt aufgerufen 01.02.2016. URL: http://www.tei-c.org/index.xml.

[248] *The Legendary Star Wars Expanded Universe Turns a New Page*. zuletzt aufgerufen 01.02.2016. StarWars.com. 2014. URL: http://www.starwars.com/news/the-legendary-star-wars-expanded-universe-turns-a-new-page.

[249] *The One Wiki to Rule Them All*. zuletzt aufgerufen 01.02.2016. URL: http://lotr.wikia.com/wiki/Main_Page.

[250] The Unicode Consortium. *The Unicode Standard*. Techn. Ber. Version 6.0.0. Mountain View, CA: Unicode Consortium, 2011. URL: http://www.unicode.org/versions/Unicode6.0.0/.

[251] Elaine G Toms. „Serendipitous Information Retrieval.“ In: *DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*. Zurich. 2000.

[252] Jeffrey Travers und Stanley Milgram. „An Experimental Study of the Small World Problem“. English. In: *Sociometry* 32.4 (1969), pp. 425–443. ISSN: 00380431.

[253] Daniel Tunkelang. *Faceted Search*. Bd. 5. Morgan & Claypool Publishers, 2009.

[254] *Understanding Shakespeare*. zuletzt aufgerufen 01.02.2016. Jstor. URL: http://labs.jstor.org/shakespeare/.

[255] *Using large-scale brain simulations for machine learning and A.I.* zuletzt aufgerufen 01.02.2016. Google. 2012. URL: https://googleblog.blogspot.de/2012/06/using-large-scale-brain-simulations-for.html.

[256] Stefania Vitali, James B. Glattfelder und Stefano Battiston. „The network of global corporate control“. In: *CoRR* abs/1107.5728 (2011).

[257] Göttingen Von Jürgen Hövermann. „Das Geographische Praktikum des Claudius Ptolemaeus (um 150 pCn) und das geographische Weltbild der Antike“. In: (1980).

[258] Denny Vrandečić. „Wikidata: A new platform for collaborative data collection“. In: *Proceedings of the 21st international conference companion on World Wide Web*. ACM. 2012, S. 1063–1064.

[259] Chris Walshaw. „A multilevel algorithm for force-directed graph drawing“. In: *Graph Drawing*. Springer. 2001, S. 171–182.

[260] D.J. Watts und S.H. Strogatz. „Collective dynamics of 'small-world' networks“. In: *Nature* 393 (1998), S. 440–442.

[261] Duncan J. Watts. *Small worlds. the dynamics of networks between order and randomness*. Princeton studies in complexity. Quelldatenbank: KUGUSBJSON ; Format:marcform: print ; Umfang: XV, 262 S. : Ill., graph. Dsrst. Signatur: 27A8351. Status: verfügbar.: Princeton Univ. Press, 1999. ISBN: 0-691-00541-9.

[262] Martin Weber. *TABVLA PEVTINGERIANA*. Dez. 2007. URL: `http://www.tabula-peutingeriana.de/`.

[263] Max Weber. *Wirtschaft und Gesellschaft*. Hrsg. von Dr. Alexander Ulfig. Zweitausendeins, 2005, S. 1138.

[264] *WikiGalaxy*. zuletzt aufgerufen 01.02.2016. URL: `http://wiki.polyfra.me/`.

[265] *Will in Town*. zuletzt aufgerufen 01.02.2016. Will in Town - Team. URL: `https://www.androidpit.de/app/de.uni_koeln.willintown`.

[266] Ian H Witten und Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[267] Christof Wolf. „Egozentrierte Netzwerke: Datenerhebung und Datenanalyse“. German. In: *Handbuch Netzwerkforschung*. Hrsg. von Christian Stegbauer und Roger Häußling. VS Verlag für Sozialwissenschaften, 2010, S. 471–483. ISBN: 978-3-531-15808-2.

[268] *Wookieepedia*. zuletzt aufgerufen 01.02.2016. URL: `http://starwars.wikia.com/wiki/Main_Page`.

[269] *WordHoard*. zuletzt aufgerufen 01.02.2016. Northwestern University. 2011. URL: `http://wordhoard.northwestern.edu/`.

[270] *Xeletor XML Server*. zuletzt aufgerufen 01.02.2016. 2014. URL: `http://sourceforge.net/projects/xeletor.berlios/`.

[271] Alon Zakai. „Emscripten: An LLVM-to-JavaScript Compiler“. In: *Proceedings of the ACM International Conference Companion on Object Oriented Programming Systems Languages and Applications Companion*. OOPSLA '11. Portland, Oregon, USA: ACM, 2011, S. 301–312. ISBN: 978-1-4503-0942-4.

[272] Thomas Zerres. *Bürgerliches Recht: eine Einführung in das Zivilrecht und die Grundzüge des Zivilprozessrechts*. Springer-Verlag, 2013.

[273] *Zgharta Liberation Army*. zuletzt aufgerufen 01.02.2016. Wikipedia, the free encyclopedia. 2014. URL: `http://en.wikipedia.org/w/index.php?title=Zgharta_Liberation_Army&oldid=593193216`.

[274] Tao Zhou u. a. „Bipartite network projection and personal recommendation“. In: *Phys. Rev. E* 76 (4 2007), S. 046115.

[275] Mengia Zollinger, Cosmin Basca und Abraham Bernstein. *Market-Based SPARQL Brokerage: Towards Economic Incentives for Linked Data Growth*. Techn. Ber. 2013, S. 282–284.

[276] Katharina Anna Zweig und Michael Kaufmann. „A systematic approach to the one-mode projection of bipartite graphs.“ In: *Social Netw. Analys. Mining* 1.3 (2011), S. 187–218.